Computational design with flexible backbone sampling for protein remodeling and scaffolding of complex binding sites

Bruno Emanuel Ferreira de Sousa Correia

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Dissertation presented to obtain the Doutoramento (Ph.D.) degree in Biochemistry at the Instituto de Tecnologia Química e Biológica da Universidade Nova de Lisboa

Oeiras, Outubro, 2010

Abstract

Computational protein design has achieved several milestones, including the design of a new protein fold, the design of enzymes for reactions that lack natural catalysts, and the re-engineering of protein-protein and protein-DNA binding specificity. These achievements have spurred demand to apply protein design methods to a wider array of research problems. However, the existing computational methods have largely relied on fixed-backbone approaches that may limit the scope of problems that can be tackled. Here, we describe four computational protocols - side chain grafting, flexible backbone remodeling, backbone grafting, and *de novo* scaffold design - that expand the methodological protein design repertoire, three of which incorporate backbone flexibility. Briefly, in the side chain grafting method, side chains of a structural motif are transplanted to a protein with a similar backbone conformation; in flexible backbone remodeling, *de novo* segments of backbone are built and designed; in backbone grafting, structural motifs are explicitly grafted onto other proteins; and in *de novo* scaffolding, a protein is folded and designed around a structural motif. We developed these new methods for the design of epitope-scaffold vaccines in which viral neutralization epitopes of known three-dimensional structure were transplanted onto non-viral scaffold proteins for conformational stabilization and immune presentation. Epitope-scaffolds were designed using the different methods for three different epitopes from Human Immunodeficiency Virus and Human Respiratory Syncytial Virus. In each case, the goal was to design stable molecules with well-defined folding transitions and high affinities for a virus-neutralizing monoclonal antibody. By experimentally testing our designed scaffolds, we demonstrated that this goal was achieved by all four methods. Further, crystallographic analysis of epitope-scaffolds designed by side-chain grafting and remodeling revealed excellent agreement between design models and experimental structures. The epitope-scaffolds designed here will serve as promising vaccine candidates. The computational design methods are generalizable and could be applied to other problems. Additionally they may be valid alternatives to problems addressed in the past with more conservative strategies.

Sumário

Os métodos computacionais existentes para o *design* de proteínas têm vindo a alcançar objectivos notáveis, tais como: o *design* de uma proteína com uma nova estrutura tridimensional; o *design* de uma enzima com actividade catalítica para a qual não são conhecidos outros catalisadores existentes na natureza e a manipulação das especificidades existentes nas interacções proteína-proteína e proteína-DNA. O alcance das metas mencionadas potenciou o interesse na aplicação de técnicas de *design* de proteínas em problemas bastante diversificados. Até então, a maioria dos métodos computacionais utilizados não permitiam qualquer tipo de flexibilidade da cadeia principal. Esta restrição na flexibilidade poderá potencialmente representar uma limitação para os problemas que podem ser abordados usando *design* computacional. Nesta tese estão descritos quatro novos protocolos computacionais: transplante de cadeias laterais, remodelação da cadeia principal, transplante de motivos estruturais e o novo *design* de proteínas-suporte. Os quatro métodos descritos diversificam o reportório metodológico existente relativamente ao design de proteínas; três deles incluem a utilização de liberdade conformacional da cadeia polipeptídica. Resumidamente descrevem-se algumas características que distinguem os diferentes métodos: para o transplante de cadeias laterais, estas são transplantadas para proteínas com conformações da cadeia principal similares às conformações do motivo estrutural de interesse; na remodelação da cadeia principal, novos segmentos da mesma são construídos com sequências de aminoácidos adequadas; no transplante de motivos estruturais, estes são transplantados ignorando a conformação local das proteínas-suporte e é permitida flexibilidade na cadeia principal adjacente para estabilização do motivo transplantado; finalmente, no *de novo design* de proteínas-suporte, a maioria da cadeia polipeptídica tem liberdade conformacional, suportando o motivo estrutural de interesse. Estas novas metodologias foram aplicadas na geração de proteínas para o suporte de epítopos; os epítopos, pertencentes a proteínas virais com a estrutura tridimensional determinada, foram transplantados. A transplantação do epítopo pretende assegurar a estabilização estrutural, para que estes sejam apresentados

ao sistema imunológico de forma eficiente. As proteínas-suporte foram criadas contendo epítopos para os quais são conhecidos anticorpos neutralizantes pertencentes a dois vírus: o Vírus de Imunodeficiência Humana e o Vírus Sincicial Respiratório. Para ambos os vírus, as proteínas criadas computacionalmente foram testadas experimentalmente demonstrando ser termicamente estáveis e possuírem elevadas afinidades de ligação com anticorpos neutralizantes. As estruturas cristalinas das proteínas criadas computacionalmente estão em concordância com os modelos computacionais inicialmente gerados. A longo prazo estas moléculas poderão ser utilizadas para o desenvolvimento de vacinas. As metodologias computacionais desenvolvidas são generalizáveis e potencialmente úteis para outras aplicações. Os protocolos desenvolvidos constituem, também, alternativas viáveis para serem aplicados em problemas de *design* computacional de proteínas que, anteriormente, foram abordados com estratégias mais conservadoras.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my supervisors Bill Schief and David Baker for hosting me in the University of Washington in Seattle, for their guidance and for such a good time learning science from them. To Andreia, my companion in life for her love and understanding throughout the good and bad moments. To my parents and family for their unconditional love and support since ever, more so in this time we have been far apart.

I would like to thank to Instituto Gulbenkian Ciência- PhD Program in Computational Biology, in particular to Jorge Carneiro e Marie-France Sagot for this lifetime opportunity that I was fortunate enough to have. To Instituto de Tecnologia Química e Biológica da Universidade Nova de Lisboa for their institutional support, in particular to my thesis committee Claudio Soares and António Baptista for their good will to take the time to give me useful advices.

To my first scientific mentors Rui Brito e Teresa Pinheiro for their ability to raise my passion for science. I would also like to thank to my laboratory colleagues by always providing the best intellectual and social environment that made everything easier. Very importantly I want to acknowledge all the collaborators which I had the chance of working with, in particular the Stamatatos and Strong Labs.

To all my friends that from childhood and college that were always supportive and kind enough to receive me with their open arms. I would like to thank to my friend Toni, until today his memory is still a source of inspiration. Finally, I would like to thank to Fundação para a Ciência e Tecnologia for providing the funding that was absolutely critical. More than dedicating this work to all this people, I feel that each and everyone of you owns an important part of it.

ABBREVIATIONS

HIV - Human Immunodeficiency Virus

RSV - Respiratory Syncytial Virus

SIV - Simian Immunodeficiency Virus

IgG - Immunoglobulin G

Fab - Fragment antigen-binding

Fv - variable fragment

CDRH - Complementarity-Determining Region on the Heavy chain

Ab - Antibody

MAb - Monoclonal Antibody

bN - broadly Neutralizing

SPR - Surface Plasmon Resonance

CD - Circular Dichroism

NMR - Nuclear Magnetic Resonance

SEC -Size Exclusion Chromatography

SLS -Static Light Scattering

SDS-PAGE - sodium dodecyl sulfate polyacrylamide gel electrophoresis

BCA - Bradford Colorimetric Assay

CCD - Cyclic-Coordinate Descent

FFL - Fold From Loops

RMSD - Root Mean Square Deviation

CD4 - Cluster of Differentiation 4

Chapter 1

GENERAL INTRODUCTION

1.1 Protein Design - Historical Perspective

A long-standing goal of a wide scientific community is to be able to create tailored proteins that efficiently perform desired biological functions. The three-dimensional structure of a protein and its interaction partners is thought to determine the function of that protein, and a major aim of structural biology is to dissect the relationship between structure and function and to understand how different structural features govern different functions.

Though predicting the structure or function of a protein from its amino acid sequence remains an unsolved problem, finding sequences compatible with a particular structure (and function) may be more tractable, as first discussed by Drexler[1]. The process of selecting sequences that will fold into a known structure was defined by Pabo[2] as the "inverse folding" problem and has been the strategy of choice to tackle most protein design problems to date. One of the first attempts to employ a flavor of the inverse folding strategy was reported by Gutte et al. in 1979[3] where a 34-residue polypeptide was designed, based on rules extracted from the few known x-ray structures. A simple topology was idealized with the purpose of binding to nucleic acids; with this back of the envelope topology they choose an amino-acid sequence with a predicted secondary structure in agreement with the target topology. After 30 years many of these principles are still employed. Remarkably, Gutte and coworkers had the conceptual view of designing a structure to accomplish a defined function, though as the field of protein design has progressed most studies have aimed lower, with purely structural aims disregarding any functional purposes.

Structural biology has come a long way since the first attempts to rationally design proteins, and the expanding Protein Data Bank[4] - currently holding approximately 70,000 protein structures - provides a wealth of structural information for the success of the inverse folding strategy. Likewise, computational structural biology has flourished, based on massive

amounts of experimental information available, active development of efficient algorithms and vastly increased computational resources.

Ponder and Richards[5] elaborated on Drexlers suggestion and devised one of the first computational methods to approach the inverse folding problem, by creating the concept of "tertiary templates". The tertiary templates consist on a set of sequences compatible with a prototype crystal structure according to specified set of stereochemical rules. To sample side chain degrees of freedom, Ponder and Richards used a "rotamer" library approach while keeping the protein backbone fixed. The computational algorithm favored combinations of amino acids in the core free of steric clash but optimized the packing density.

Advances in protein computational design were also intimately related with advances in molecular biology and DNA manipulation, which allowed for routine experimental characterization of designed proteins. Hellinga and Richards, also pioneering protein designers, described in 1991 the DEZYMER [6], a computer program able to search for constellations of backbone positions in a protein to design ligand binding sites according to a predefined geometry of interaction between ligand and protein. The design of a metal binding site in thioredoxin was accomplished with DEZYMER, being one of the first successes of computational-aided design[7].

1.2 Milestone Achievements in Computational Protein Design

1.2.1 Protein Stability Enhancement

After the first groundbreaking steps in computational protein design, many milestones envisioned by Drexler and Pabo have been achieved. One of the first applications of protein computational design was the enhancement of protein stability, a central property for the development of molecules with biological and therapeutic applications. In general, improvements have been achieved by increasing the hydrophobic surface burial, in most of the cases more hydrophobic residues in the redesigned core translate to improved protein stability[8]. It is worth mentioning that substitutions of polar amino acids by hydrophobic are not always advantageous as showed in work from the Mayo laboratory[9, 10] , where by counting the hydrogen bonds made by a particular residue, it was possible to identify which residues were

important to keep intact. The same group demonstrated that other electrostatic principles are also efficient approaches to improve protein stability, such as mutations that stabilize the helix dipole and helix capping mutations[9, 10, 11].

1.2.2 Design of existing and novel protein folds

The complete re-design of existing folds or the design of a novel topology are strict tests of the performance of energy functions and of our understanding of the physical chemistry governing protein structure. The large number of sequences possible increases the demand for energy functions able to discriminate which sequences are energetically more favorable for a given fold. The test becomes even more stringent if we consider that computationally designed proteins frequently have the correct secondary structure but often fail to attain the desired tertiary structure, behaving as molten globules or partially unfolded protein with tendency for aggregation. Dahiyat et al. described one of the first designs of a whole protein structure in which the target topology was a 34 residue $\beta\beta\alpha$ zinc finger motif [12]. The algorithm relied on the dead-end elimination theorem for a rapid and efficient selection of sequences and rotamers [13]. Different scoring functions were used for positions within the protein core, surface, or a "boundary" region between the core and surface. Core positions were designed allowing only hydrophobic amino acids, surface positions allowing only hydrophilic, and boundary positions allowing both types of amino acids. The final designed sequence had minimal sequence similarity to any zinc-finger sequence known by that time. Interestingly, designed hydrophobic residues in the protein core replaced a native metal binding site. The agreement between the computational model and determined solution structure was 1.98 Å RMSD over the backbone, and similar core packing was observed in both the model and solution structure.

The design of a new protein fold poses a more stringent test to the computational algorithms because there is no pre-existing backbone template that may carry some bias to native sequences [14, 15, 16, 17, 18]. Harbury et al. [19] designed a family of dimeric, trimeric and tetrameric coiled-coils with pre-determined right-handed super-helical fold. This seminal work explored structural diversity of the coiled-coils using algebraic parameterization,

originally described by Francis Crick[20]. Right-handed coiled-coils were generated and designed following an undecatad amino acid repeat in which 3 of the positions were restricted to hydrophobic amino acids. A structure of a designed tetrameric coiled coil was solved and showed a backbone RMSD of 0.2 Å compared to the computational model. Despite the successful results, an obvious limitation is that this approach cannot be readily applied to more complicated topologies that lack defined parametric descriptions. Kuhlman et al. described the design of a novel protein fold (Top7)[21]. The authors devised a set of distance constraints that defined a topology not yet observed in nature and employed Rosetta *de novo* structure prediction[22] to generate the initial models for design (further described in section 1.3). Through iterations between sequence design and structure optimization, low energy sequence-structure pairs were found. Core positions were restricted to hydrophobic amino acids while surface exposed were restricted to hydrophilic. The computational model and the crystal structure of the design showed an RMSD of 1.2 Å and Top7 showed high thermal stability. Top7 was unprecedented by showing that it is possible to sketch a protein fold not observed in nature and design it to became a real protein. It was also unprecedented in the use of backbone flexibility both to build initial design templates and to iteratively search for low energy sequence-structure pairs (the use of flexibility will be further discuss in section 1.4).

1.2.3 Interface Design

Beyond designing for structural recapitulation, the ultimate goal of protein engineers is to be able to create functional molecules with biotechnological and therapeutic applications, or research value where they can serve as probes to help the understanding of cellular processes. Protein-protein interactions are central to many processes in cells, and the design and manipulation of protein-protein interactions has been the subject and motivation for many developments in protein computational design. As clearly stated by Karanicolas and Kuhlman[23], interface design objectives typically fall into one of three categories: affinity enhancement of naturally occurring interfaces, specificity switches in protein interactions, and designing interactions from scratch which is yet to be attained.

Tidor and co-workers reported a 140-fold improvement in affinity between an antibody-antigen complex, achieved by searching for mutations with favorable electrostatic interactions. Additionally, they showed better correlation in predicting affinity enhancing mutations by only considering electrostatics interactions opposed to the interaction energy. From a collection of other studies the decrease of the desolvation cost of the interface and the increase of hydrophobic surface area buried in the interface seem like the most efficient way to improve affinities between complexes[23].

Several successes have been reported in designing interface specificity. Multistate design was one of the successful strategies, in which positive design favoring one ore more target interactions and negative design disfavoring competing undesired interactions are carried out simultaneously. Using such a strategy Havranek and Harbury reported the design of selective coiled coil pairs [24]. More recently, Keating and co-workers designed specific peptide partners for 19 of 20 human basic-region leucine zipper (bZIP) transcription factors. The synthetic peptides were specific for the target family and not for the remaining 18 families. For fast computation of energies to the desired targets and competitor targets it was used a cluster expansion method that allows for the conversion of structure-based model into a sequence-base scoring function[25]. The cluster expansion method was trained on a set of sequences that were modeled structurally. bZIP peptides are a fairly structurally conserved family, so the question remains if such methods will be applicable to families structurally more diverse[23].

According to the literature the use of multistate design seems to be required if the degree of similarity of the competitors is high. Multistate design has also been used in a scheme where the desired target is the interface to be designed and the negative target is the protein in the absence of the binding partner. This scheme improves sequence recovery in interface design (S. Fleishman personal communication) where the existence of a competing state may compensate for inaccuracies in the energy function.

In the absence of appreciable similarity between competing states, strategies based in positive design seem to be efficient and have been employed to modulate calmodulin binding specificity. Shifman and Mayo were able to maintain the affinity for the target interaction (calmodulin - smMLCK) while reducing the affinity for other 6 known interacting peptides

by 1.5 to 86 fold[26]. Soon after, Reina et al. reported the engineering of PDZ domains to recognize new targets using a conceptually similar strategy[27].

Another important achievement in the design of interfaces of macromolecules was the rational redesign of protein-DNA interactions to modulate binding and cleavage specificity of a homo-endonuclease (Mso). Base-pair substitutions were performed and clusters of residues responsible by accommodating the newly introduced base pair were redesigned. The specificity switch achieved for the new DNA sequence relative to the wild type was 10000 fold, and the computational design showed good agreement with determined crystal structure [28].

1.2.4 De novo Enzyme Design

De novo design of catalytic activities not observed in natural enzymes has been another major goal of protein design since the early days when Drexler mentioned the possibility of designing molecular machines[1]. Hellinga and Richards contributed critically with the idea of searching constellations of amino acids in which functional groups could be placed in a pre-defined geometry[7]. Later on, matching algorithms using geometrical hashing were developed by multiple groups to identify structurally suitable scaffold proteins for enzymatic sites[29]. Baker and co-workers reported the first enzymes that catalyze reactions not observed in nature by employing a hashing algorithm to find positions to host catalytic motifs that had been optimized by quantum mechanical calculations[30, 31, 32]. In both cases the catalytic activity of the computational designs was far less than the activities observed in natural enzymes, nevertheless these designs served has a proof of principle, posed stringent tests to the understanding of enzyme catalysis, and also provided valuable starting points for directed evolution.

1.2.5 Scaffolding with Rational Approaches

The transplantation of side chains, structural motifs, binding sites or active sites from native proteins to other non-related proteins can be generally categorized as scaffolding. Typically for these design problems there is a first stage to select host proteins (scaffolds), a

second stage of transplantation of the motif of interest, and a third stage to design first and second shell residues in order to find energetically favorable sequences of amino acids. In addition to DEZYMER, several other computational algorithms have been reported, that find appropriate geometries to graft side chains of linear or non linear sequences[33, 34, 35]. At least two reports have described side chain grafting of helical motifs to the transcription factor GCN4 Serrano and colleagues transplanted part of an interleukin-4 binding site[36] and Kim and colleagues grafted a region of HIV gp41[37]. Both of these grafting exercises were successful, in the sense that several designed molecules showed high affinities to the intended targets, nevertheless both skipped a systematic search for accepting scaffolds basing the scaffold selection in secondary structure similarity and both relied on the alpha-helical geometry. A more systematic approach for scaffolding was taken in Liu et al.[35] in which a protein-protein interaction was created by grafting the key residues of one of the wild type proteins onto a non-related protein. In this study, three residues were grafted onto a scaffold identified by a geometrical hashing search of the PDB; the K_Ds to the target binding site were 100 fold higher in comparison to the wild type protein. The computational design was not confirmed by experimental structural characterization, nevertheless functional studies *in vivo* showed the expected activity.

1.3 Rosetta - The Tool

The Rosetta software package was initially created to perform *ab initio* protein prediction [22]. Since then, Rosetta has been a fast-developing platform with a myriad of applications regarding to macromolecular modeling and design. Besides protein structure prediction, Rosetta has acquired many other functionalities such as loop modeling[38], protein docking[39] and small molecule docking[40], protein design, phasing diffraction data with *de novo* models[38], usage of NMR[41] and cryoelectron microscopy[42] data for structure refinement, DNA-protein interface design[28, 43] and RNA folding[44], among others.

The core features of Rosetta that enable this variety of modeling applications are low and high-resolution energy functions, a variety of conformational sampling methods, and use of both Metropolis Monte Carlo criteria and minimization algorithms to reach energy minima.

In stages where a large conformational space is sampled, a coarse-grained low resolution potential is often used first, together with a reduced description of the protein structure where side chains are represented as centroids. The low resolution energy function is dominated by several statistical potentials such as solvation and electrostatics derived from observed residue distributions on protein structures. Explicit hydrogen bonds are not considered but probabilistic terms to evaluate secondary structure motifs are taken into account. Steric overlaps between the side chain centroids and backbone atoms are penalized but overall compact structures are favored by computing van der Waals interactions and gyration radius of the protein. After a first coarse-grain exploration of the energetic landscape, typically several local minima are found; to search for the lowest energy minimum all the atomic detail is added to the protein structures, allowing a more accurate calculation of the energetic contributions. The full-atom energy function computes the interactions between buried atoms in an attempt to capture the high energetic cost of void volumes in the protein core, van der Waals interactions are modeled with a Lennard-Jones potential. The solvation is included using the implicit solvation model of Lazaridis and Karplus [45] and the energy for explicit hydrogen bonds is computed using a secondary structure and orientation dependent potential[46]. In terms of conformational sampling strategies there is a wide variety of moves that can be applied to the polypeptide chain depending on the atomic detail of the sampling stage and also of the degree of conformational diversity desired. Generally, at a low resolution stages of sampling, moves that yield large structural changes are needed, fragment insertion is typically the first stage of conformational sampling for standard protocols as *de novo* structure prediction and loop prediction. Fragment libraries are constructed from short fragments of known protein structures, with a residue window of length 3 and 9, and are selected according sequence identity, secondary structure or a mixture of both. A fragment insertion replaces the torsion angles of the polypeptide chain with the torsion angles of the fragment. After the fragment insertion stage several types of moves are employed with a finer sampling scope: random torsion angle perturbation, selection of globally non-perturbing fragments and torsion angle variation to offset global perturbations of the structure. At high resolution, to further sample the local conformational landscape the geometries are optimized by multiple torsion angle perturbations

followed by rotamer optimization and continuous gradient based minimization of side chain and backbone torsion angles. The main focus of the work presented here is protein design, it is worth mention that RosettaDesign[15] shares the same principles described for the high resolution energy function. Different amino acids are considered in each sequence position and side-chain conformation included in a rotamer library are sampled[47]. A Metropolis Monte Carlo procedure is used to accept or reject the move that consists in exchanging one rotamer for another at a randomly chosen position.

Currently, Rosetta is a flexible platform for molecular modeling providing scoring functions and conformational sampling moves suited for a wide range of resolutions. Due to its modular architecture, data derived from a variety of experimental techniques can be easily used to guide and validate computational calculations.

1.4 Fixed versus Flexible Backbone Design

As briefly reviewed here, many successes using computational protein design have been reported, however only a few notable exceptions allowed backbone flexibility in conjunction with design calculations. The prospect for generalized use of flexibility in computational protein design holds the promise of meeting the challenges for the next generation of protein engineering. It is reasonable to expect that flexible backbone modeling will open new perspectives: allow the design of more efficient enzymes by trying to model more precise geometries for catalytic machineries; design of conformational and specificity switches; and fundamental understanding of the relationship between function and dynamics in proteins.

Utilizing an existing structure to be designed with a fixed backbone approach guarantees that such conformation is possible and the design problem is reduced to find the combination of amino acids that best fit the target conformation. On the other hand, if the design template has never been observed the problem becomes more complex because there are no guarantees that the template conformation is realistic.

Besides the milestone designs by Harbury et al. and Kuhlman et al., mentioned above, backbone flexibility has been used to design a loop with sub-angstrom accuracy[48] and to switch the substrate specificity of an enzyme[49]. Backbone flexibility in the first report was achieved by small perturbations of the backbone and in the second by a fragment insertion

strategy.

Given the promise and current limitations of flexible backbone protein design, it is important to pursue new methods and extensions of the existing methods . Several important questions on flexible backbone design will be addressed in this thesis: i) How realistic is the backbone conformational space sampled?; ii) Can we productively bias the conformational sampling by adding functional and structural constraints? ; iii) How to discriminate good from bad designs?; iv) Can we inform high-throughput selection techniques using computational design ? By providing some solutions to these problems, it is hoped that we have advanced the field of protein design and that others will build on the work presented here.

1.5 Biological Relevance and Thesis Overview

1.5.1 Biological Relevance

The current structural knowledge regarding pathogens and other disease-related targets provides many opportunities to devise structure-based strategies with potential therapeutic applications. The transplantation of broadly neutralizing epitopes sets itself as a tempting strategy to design immunogens to re-elicit neutralizing antibodies, and consequently being useful as vaccine candidates, so called retrovaccinology approach[50]. Current successful vaccines contain either the entire inactivated pathogen or isolated molecules known to generate protective immunity. Despite the considerable success of traditional empirical vaccine design approaches, many pathogens have eluded vaccine design with these approaches. Two such viruses are used as case studies in this work - HIV (Human Immunodeficiency Virus) and RSV (Respiratory Syncytial Virus). Native antigen molecules are adapted to escape rather than induce immunity, and undesired properties for efficient induction of neutralizing responses are observed in the native antigens. In principle these undesired properties can be optimized by rational methodologies if structural knowledge is available. Examples of such optimization include: trimming variable regions, reducing multi-domain proteins to minimal domains that contain the neutralizing epitopes, and stabilizing the conformation of neutralizing epitopes. Many of these potential improvements are directly related to an important challenge for immunogen design and fundamental immunology, that is to under-

stand and engineer immunogenicity and learn how to make an epitope more "obvious" to the immune system. Rational approaches can be devised and tested to overcome the lack of knowledge about the principles that govern immunogenicity, in an attempt to conceptualize a set of practical strategies that will allow to improve immunogenic properties.

Considering all the hurdles with traditional approaches to generate useful immunogens for pathogens that still lack vaccine candidates, we sought to use rational methodologies to transplant broadly neutralizing epitopes to non-related protein scaffolds. A successful transplantation could overcome several of the culprits identified on native molecules - for example this strategy will allow the presentation of epitopes that are only transiently exposed on the pathogen, for example it is thought that the 4E10 epitope on HIV gp41 is only available on the virus-cell fusion stage[51]. Additionally, a successful transplantation would also allow for precise stabilization of the epitope conformation recognized by the broadly neutralizing antibodies - hence scaffolding is a promising rational strategy to try to re-elicit broadly neutralizing antibodies.

At the beginning of my project the state of art of both immunogen and computational design was rather different. Reports of rational and structure-based approaches applied to immunogen design were sparse; the aims set for my PhD were to develop methods that would allow dealing with a variety of design challenges. In the process of developing the design methodologies we aimed to generate useful reagents for the vaccine research community. Methodologically, approaches using flexible backbone design had only been described in few reports [19, 15]. Due to the nature of the design problems we were facing, we aimed to develop methodologies to routinely incorporate flexible backbone in computational design calculations.

1.5.2 Thesis Overview

In my project I developed and applied computational techniques to transplant broadly neutralizing epitopes from HIV and RSV to carrier proteins, the epitope-scaffolds. Due to the structural diversity and complexity of different epitopes, several computational methods had to be developed to successfully design epitope-scaffolds.Chapter 2 describes a fixed backbone

design method, side chain grafting, that was developed to transplant linear epitopes. The matching stage for selection of acceptor protein scaffolds relied mostly in two criteria: backbone similarity between the acceptor scaffold and the epitope of interest; and structural compatibility between the acceptor scaffold and the antibody binding orientation. Computational designs were structurally and biochemically characterized, and immunogenicity studies were performed.

A natural continuation ion to the design of the first generation epitope-scaffolds and also following-up the results from immunogenicity studies from the previous chapter, in Chapter 3 describes two strategies to improve the immunogenic properties of the epitope-scaffolds. Two strategies were attempted to improve the immunogenic response to the epitope of interest: domain trimming and resurfacing. For the domain trimming, I devised a flexible backbone design protocol to rebuilt and design *de novo* segments with pre-defined secondary structure. Another rational approach described is resurfacing that uses fixed backbone design to change the majority of the epitope-scaffold surface residues creating panels of molecules where the common trait is the epitope. Experimental results validate the novel flexible backbone design and also show desirable biophysical properties of the resurfaced designs.

Most of the neutralizing epitopes are not contiguous, instead they are defined by several discontinuous segments of the protein that interacts with the antibody. As consequence, the method presented in Chapter 2 was inappropriate to deal with complex structural motifs. Chapter 4 describes the development of a new method to transplant complex epitopes. The flexible backbone philosophy described in Chapter 3 was extended to the modeling of several segments in a protein such that the complex epitope graft would be structurally supported. Due to the complexity of the design problem, computationally based directed libraries were designed and *in vitro* evolution was leveraged by the computational protocol, providing a fast and efficient protocol for screening and selection of sequences with improved binding affinity to the target antibody.

Most scaffolding strategies are inherently limited to the existence of scaffolds compatible with the epitope of interest. To further expand the modeling strategies available for epitope transplantation, in Chapter 5 I developed a computational method for *"de novo"* design of

epitope-scaffolds. This method only requires a target topology compatible with the desired epitope, dismissing the atomic detail considered by the matching algorithms described in the previous chapters. In principle, it is not required that the target topology has been previously observed in nature, any "back-of-the-envelop" topology should be sufficient to serve as input for the protocol. The computational methodology is conceptually different from the previous protocol in which a structural motif is forced in a mostly rigid scaffold. In contrast, in the new *"de novo"* approach, only the epitope is rigid and the rest of the scaffold is moveable in order to achieve the maximal support for the rigid segment. To evaluate the performance of the method computational designs were experimentally characterized. It is important to reinforce that the methods developed throughout my PhD project were applied to epitope scaffolding and immunogen design, nevertheless they are generalizable and readily applicable to other biological problems such as the design of protein inhibitors and enzymes.

Chapter 2

COMPUTATIONAL DESIGN OF EPITOPE-SCAFFOLDS ALLOWS INDUCTION OF ANTIBODIES SPECIFIC FOR A POORLY IMMUNOGENIC HIV VACCINE EPITOPE

2.1 Abstract

Broadly cross-reactive monoclonal antibodies define epitopes for vaccine development against HIV and other highly mutable viruses. Crystal structures are available for several such antibody-epitope complexes, but methods are needed to translate that structural information into immunogens that re-elicit similar antibodies. We describe a general computational method to design epitope-scaffolds in which contiguous structural epitopes are transplanted to scaffold proteins for conformational stabilization and immune presentation. Epitope-scaffolds designed for the poorly immunogenic but conserved HIV epitope 4E10 exhibited high epitope structural mimicry, bound with higher affinities to monoclonal antibody (MAb) 4E10 than the cognate peptide, and inhibited HIV-neutralization by HIV+ sera. Rabbit immunization with an epitope-scaffold induced antibodies with structural specificity highly similar to MAb 4E10, an important advance toward elicitation of neutralizing activity. The results demonstrate that computationally designed epitope-scaffolds are valuable as structure-specific serological reagents and as immunogens to elicit antibodies with pre-determined structural specificity.

2.2 Introduction

A protective HIV vaccine, or a universal flu vaccine, likely will elicit antibodies that block infection of diverse circulating strains of HIV or Influenza. Several anti-HIV antibodies with broad neutralizing reactivities have been isolated from natural infection, and approximately 25% of HIV-1-infected individuals produce broadly-neutralizing sera [52], but efforts to elicit such responses by vaccination have failed. Anti-influenza broadly neutralizing antibodies (bNAbs) also have been characterized [53, 54, 55, 56], but the annual influenza vaccine elicits

potent neutralizing antibodies only against the vaccine strains and closely-related isolates.

The anti-HIV MAb 4E10 neutralizes 98% of HIV strains [57] hence elicitation of 4E10-like antibodies is a major goal for HIV vaccine design. 4E10 binds a contiguous epitope in the Membrane Proximal External Region (MPER) of gp41 just N-terminal to the transmembrane domain[58, 59], an epitope that is likely only transiently exposed during viral fusion with the target cell [51]. Nuclear magnetic resonance structural analysis of MPER peptides in lipid micelles found a kinked helix partially embedded in lipid [60]. Crystal structures of MAb 4E10 in complex with gp41 peptides revealed that 4E10 binds to a helical epitope in an end-on manner with the N-terminus and one face of the helix docked into a hydrophobic cavity on the antibody; the bound epitope adopts a short 3/10 helix followed by an alpha helix, and conserved HIV side chains from both regions make significant antibody contacts [61, 62, 63, 64]. The antibody also has a long, hydrophobic complementarity-determining region H3 (CDRH3) loop, which does not bind directly to the helical epitope but which may provide contacts to the viral lipid membrane and contribute to neutralization [61, 62, 63, 64]. Here, as a first step to elicit antibodies with the binding and neutralizing properties of MAb 4E10, we aimed to develop immunogens that would elicit antibodies with the exact binding specificity of MAb 4E10 for its core peptide epitope.

A major hurdle confronting 4E10-vaccine design is that the epitope is naturally poorly immunogenic. While 4E10-like activity has been detected in HIV+ patients with broadly-neutralizing activity, the frequency is low ($\sim$ 5%)[57, 65, 66]. Furthermore, previous attempts to elicit 4E10-like antibodies by vaccination using soluble trimeric gp140[67], VLP-membrane-anchored gp41 variants[68], or a gp120 construct with the 4E10 epitope embedded in the highly immunogenic V2 loop [69], failed to elicit reactivity against the epitope primary sequence. MPER peptides anchored to liposomes or hepatitis B surface antigen particles have elicited modest responses against 4E10 peptides [70, 71], but whether these responses were directed to the 4E10 epitope or other epitopes was not determined. In none of these cases was the helical epitope conformation necessarily stabilized, though in all cases the antigens did bind to MAb 4E10.

Crystallographic analysis of several anti-HIV [72, 62, 73, 74] and anti-Influenza[75, 55] bNAb/epitope complexes, combined with advances in computational protein design [21],

provide new opportunities for structure-based vaccine design [76]. Here we pursued a strategy of stabilizing an epitope in its antibody-bound conformation, and presenting that stabilized epitope in an immunogenic format, in order to elicit antibodies with structural specificity similar to the target antibody. Specifically, we engineered epitope-scaffold immunogens in which the HIV 4E10 epitope was transplanted by computational design to small, non-viral scaffold proteins for conformational stabilization and presentation to the immune system without the immunodominant variable regions typical of persistent pathogens. The computational design methods are general and should be applicable for structural stabilization of a variety of contiguous epitopes of different conformations from HIV, Influenza and other viruses

2.3 Materials and Methods

2.3.1 Computational design of epitope-scaffolds

The automated matching and design stages of the side-chain grafting protocol were implemented in the Rosetta [77] modeling platform. The matching stage exhaustively searched 13337 single chains from the PDB [4] for backbone similarity with the 4E10 epitope by computing RMSD (Root Mean Square Deviation) between residues 671-680 or 672-680 in PDB ID:2FX7 [61] and all possible protein segments using matched-width sliding windows in the query scaffolds. All hits with RMSD < 1 Å (or RMSD/length < 0.1) were retained if backbone clash with antibody was less than 100 energy units. Epitope side-chains were transferred to all-glycine versions of the remaining hits, and then interfacial side-chain torsions and the rigid-body orientation of antibody and epitope-scaffold were optimized by minimization [78]. Highly conserved epitope side-chains W672, F673, I675, L679 and W680 were transplanted to all scaffolds. At moderately variable epitope positions (671, 674, 676), the most energetically favorable of the common HIV side-chains was generally transplanted (N, S, or G at 671; D, S, or N at 674; S or T at 676). Native scaffold side-chain rotamers outside the epitope were recovered, and residues near (heavy atom distance < 4 Å) the epitope or the antibody were designed, categorized as 'intra' and 'inter' positions respectively. Inter residues were allowed to be ALA, GLY, SER, THR and intra positions were

allowed to be all amino acids except CYS. Epitope-scaffolds were ranked by antibody binding energy and human-guided design was performed on the best to: (a) revert unnecessary or potentially destabilizing mutations; (b) eliminate unpaired cysteines and undesired functional sites; (c) trim extraneous domains; (d) optimize solubility. In two cases (T93 and T278) more advanced flexible backbone design was performed for domain trimming (T93) or fusion to Maltose Binding Protein (MBP) for stabilization (T278); discussion of these methods is beyond the scope of this manuscript and will be reported elsewhere (Correia et al., submitted JMB).

2.3.2 Protein expression, purification and characterization

DNA segments encoding epitope-scaffold constructs were synthesized with optimized codon usage and RNA structure (Codon Devices, Genscript Corp.), subcloned into pET29 (EMD Biosciences) and transformed into BL-21 Star *E. coli* (Invitrogen). Single colonies were expanded from starter cultures in Luria Broth into ZYP-5052 auto-induction media [79] plus Kanamycin (100 mg/mL), incubated for four hours at 37°C and then overnight at 18°C. Cultures were then pelleted and stored at -20°C. Pellets were resuspended in standard buffer (50 mM Tris, 500 mM NaCl, 10 mM Imidazole, 0.5 mg/mL lysozyme), sonicated on ice and clarified by centrifugation. Supernatants were tumbled with 10 mL of nickel-NTA resin (Superflow NTA, Qiagen) for 30 min at 4°C. The resin was then rinsed twice with 10 mM imidazole, once with standard buffer plus 20 mM imidazole and eluted with standard buffer plus 250 mM imidazole. Eluates were concentrated by ultrafiltration (Amicon Ultra, Millipore) and filtered through 0.22 micron Ultrafree-MC spin columns (Millipore). Epitope-scaffolds were then purified by preparative size exclusion chromatography (SEC) on Superdex 75 16/60 columns (GE Healthcare) at room temperature in 25 mM PIPES (pH = 7.0), 150 mM NaCl, 1 mM EDTA and 0.02% w/w sodium azide (PNEA). Epitope-scaffold purity was confirmed by SDS-PAGE and analytical SEC; analytical SEC ± target NAb (Superdex 75 or 200 10/30 columns, GE Healthcare) in PNEA was used to assay qualitative binding, multimer state, monodispersivity and complex stoichiometries. Protein concentrations were determined with the BCA protein assay (Pierce); solution thermostabilities (T_m)

were determined by circular dichroism (CD) spectroscopy with Jasco J-815 instrumentation as previously described [64].

In order to prepare endotoxin-free protein for immunization studies, bacterial pellets were alternately resuspended in detergent buffer (50mM NaH2PO4, 500mM NaCl, 10mM imidazole, 0.5mg/mL lysozyme, 0.01mg/mL DNAse, 0.1% Triton X114) and loaded IMAC resin was alternately initially washed in 10 mM imidazole, 50mM NaH2PO4, 500mM NaCl, 0.1% Triton X114. Preparative SEC purification was carried out in sterile PBS on columns just previously cleaned with a bolus of 1M NaOH. Endotoxin levels were assayed with the PyroGene Recombinant Factor C Endotoxin Detection System (Lonza). Protein preparations were further purified using the ProteoSpin Endotoxin Removal Maxi Kit (Norgen) if endotoxin levels exceeded 2EU/50 μg at a [protein] of 1 mg/ml, the maximum level accepted for immunization.

2.3.3 Crystallization and crystallography

Crystals of epitope-scaffolds and epitope-scaffold-Fv complexes were grown by vapor diffusion, under the conditions shown in Table 2.1. The complexes were purified as such by SEC prior to crystallization. Diffraction data were collected at -170 °C on cryopreserved crystals, as detailed in Table 2.2, and processed with d*TREK (Pflugrath, 1999). Initial structure factor phases were determined by molecular replacement (search models in Table S3), using the program Phaser [80] or MolRep [81], as implemented under CCP4i [82]. Successive rounds of modeling and positional and individual B factor refinement were carried out with Coot [83] and Refmac5 [84]. Five of the six structures were submitted to the TLSMD server [85, 86] where TLS groups were defined. Structure validation was carried out with Procheck [87], the MolProbity server [88] and the RCSB ADIT validation server. The structures were deposited in the PDB [4] with accession codes in Table 2.1. Data collection and structure refinement statistics are in Table 2.2.

Table 2.1: Crystallization conditions and structure determination.

Epitope-scaffold	**T18**	**T117**	**T161**	**T246**	**T88-Fv**	**T93-Fv**
Crystallization conditions ([protein])	0.8 M Na, K tartrate, 0.1 M Tris, pH 8.5 (11.2 mg/mL)	2.2 M di-ammonium phosphate, 0.1 M Tris, pH 8.5 (10.2 mg/mL)	0.2 M $MgCl_2$, 33% PEG 4000, 0.1 M Tris, pH 8.5 (12 mg/mL)	2.2 M ammonium sulfate, 0.1 M bis Tris, pH 6.5 (11.5 mg/mL)	0.8 M sodium acetate, 0.1 M imidazole, pH 8.0 (11.9 mg/mL)	26% PEG 2000 MME, 0.1 M Tris, pH 8.0 (12 mg/mL)
Cryoprotectant	ethylene glycol	paraffin oil	*none needed*	ethylene glycol	ethylene glycol	ethylene glycol
Detector and optics	R-AXIS IV++ w/ HR optics	R-AXIS IV++ w/ HR optics	Saturn 944+ CCD detector w/ HF optics	Saturn 944+ CCD detector w/ HF optics	Saturn 944+ CCD detector w/ HF optics	Saturn 944+ CCD detector w/ HF optics
$\mathbf{d_{min}}$	2.3 Å	1.9 Å	2.0 Å	2.5 Å	2.65 Å	2.8 Å
Molecular replacement search model	computationally-designed model with the epitope removed	computationally-designed model with the epitope removed	computationally-designed model with the epitope removed	computationally-designed model with the epitope removed	computationally-designed complex model with the epitope removed	separate models for Fv* and computationally-designed epitope-scaffold model with the epitope removed
Molecular replacement program	Phaser	Phaser	MolRep	Phaser	Phaser	Phaser
Number of TLS groups	2	3 per chain; 6 total	4	5 per chain; 15 total	0	1-3 per chain; 8 total
PDB ID	3LEF	3LF6	3LF9	3LG7	3LH2	3LHP

Table 2.2: X-ray diffraction data and refinement statistics for the epitope-scaffold structures.

	T18	T88-Fv	T93-Fv	T117	T161	T246
Data Collection						
Space group	$P6_522$	$P2_1$	$P2_12_12_1$	$P2_12_12_1$	$C222_1$	$P2_12_12_1$
Cell dimensions	56.4, 56.4, 324.1 Å	75.9, 145.9, 78.6 Å; 92.5 °	53.0, 88.6, 150.0 Å	49.2, 82.0, 94.2 Å	40.7, 75.4, 79.6 Å	91.6, 94.9, 108.4 Å
Resolution (Å)[1]	48.30-2.30 (2.38-2.30)	28.01-2.65 (2.74-2.65)	29.54-2.70 (2.80-2.70)	49.21-1.90 (1.97-1.90)	32.64-2.00 (2.07-2.00)	28.82-2.50 (2.59-2.50)
Number observed reflections[1]	115,705 (8910)	190,285 (17,961)	66,868 (4874)	133,236 (12,691)	38,309 (2008)	150,445 (12,033)
Number unique reflections[1]	13,770 (1304)	49,017 (4883)	19,356 (1705)	30,053 (2886)	8462 (755)	33,220 (3180)
Redundancy[1]	8.40 (6.83)	3.88 (3.68)	3.45 (2.86)	4.43 (4.40)	4.53 (2.66)	4.53 (3.78)
Completeness (%)[1]	93.7(91.5)	98.9 (98.6)	96.2 (85.5)	97.5 (95.2)	98.3 (89.9)	99.5 (96.8)
Rmerge[1]	0.086 (0.373)	0.095 (0.376)	0.102 (0.405)	0.057 (0.341)	0.052 (0.185)	0.107 (0.429)
Average I/σ (I)[1]	9.0 (3.9)	9.4 (2.7)	7.1 (2.2)	17.2 (2.6)	18.6 (4.0)	6.6 (2.0)
Structure Refinement						
Resolution (Å)	46.78-2.30	26.70-2.65	28.99-2.70	31.50-1.90	27.37-2.00	28.82-2.50
R_{work}/R_{free}	0.258/0.290	0.222/0.270	0.243/0.335	0.223/0.279	0.204/0.260	0.257/0.301
Number of atoms:						
Protein	1275	9019	5112	2494	925	3155
Water	55	132	82	262	127	94
Other	4 (1 ethylene glycol)	-	4 (1 ethylene glycol)	15 (3 phosphates)	-	55 (11 sulfates)
r.m.s. deviations from ideal values:						
bond lengths (Å)	0.009	0.006	0.009	0.009	0.011	0.014
bond angles (°)	1.146	0.927	1.132	1.056	1.004	1.189
chiral volume ($Å^3$)	0.070	0.054	0.051	0.061	0.057	0.062
Ramachandran plot statistics:						
most favored regions (%)	95.8	90.4	90.0	94.1	100.0	99.4
additional allowed regions (%)	3.5	9.2	9.6	5.9	0.0	0.6
generously allowed regions (%)	0.0	0.0	0.0	0.0	0.0	0.0
disallowed regions (%)	0.7 (1 res.[2])	0.4 (4 res.[3])	0.3 (2 res.[3])	0.0	0.0	0.0
Estimated coordinate error (maximum likelihood e.s.u.) (Å)	0.19	0.240		0.17	0.137	0.208
Average B factor:						
Protein	66.4	57.2	68.1	43.6	45.1	76.6
Water	58.6	49.7	49.6	49.7	50.9	73.7
Other	85.3		53.2	82.4		96.3

[1] Values in parentheses correspond to the highest resolution shell.

[2] Residue 129, which is in a γ turn conformation, as is the 1.5 Å resolution parent structure 1Z6N.

[3] Fv light chain residue 52 (51 by Kabat numbering) of all four chains, which is in a highly conserved γ-turn conformation.

2.3.4 Sera Analysis and Imunizations

Luminex Multiplex Assay

4E10 peptides with three lysines added to their C-termini were amine-coupled to Bio-plex beads (Biorad, CA) (12μg peptide/100μl beads) as described by the manufacturers protocol. Serially diluted sera (pre- and post-immunization) or MAb 4E10 were incubated with the bead cocktail for 1hr at RT. After four washes, beads were incubated with anti-rabbit or human PE-conjugated secondary Ab for 1 hr at RT. A final wash step was performed prior to analysis on a Luminex 200 System (Invitrogen, CA). Relative avidity was computed by subtracting peptide-matched, pre-bleed mean fluorescent intensity (MFI) from each sample and then normalizing to the experiment-matched MFI for bNAb 4E10 binding to wild-type 4E10 peptide.

Peptide ELISA

Binding of rabbit sera or bNAb 4E10 to 4E10 peptide (NWFDITNWLWYIRKKK [67]), alanine-mutant 4E10 peptides, and epitope-scaffolds was assessed by ELISA as previously described [67].

Epitope scaffold competitive ELISA

Rabbit sera (1:25) were incubated with epitope-scaffolds (100 μg/mL) for 1 hr, and then the mixture was added to ELISA plates coated with recombinant gp41 (HxB2) (0.5 μg/mL) (Meridian Life Science) for 90 min at 37 ^{o}C. The binding of antibodies to gp41 was detected as above for the peptide ELISA.

Competing the neutralizing activity of HIV-1+ sera by epitope-scaffolds

Epitope-scaffolds were tested for their ability to inhibit the neutralizing activity of plasma VC10028 as previously described for peptide inhibition [66].

Immunizations

Rabbits were immunized monthly at the Pocono Rabbit Farm and Laboratory Inc (Canadensis, PA) with epitope-scaffolds mixed in PEI (Polysciences, Inc) or QS21 (Antigenics, Inc); 0.2 mg protein in 0.2 mL of PBS was administered into the two hind legs (0.2 mL per injection site). KLH-peptide immunizations were performed by GenScript Corp (Piscataway, NJ); rabbits were immunized with 0.5mg KLH-peptide and Complete Freunds Adjuvant, and boosted at weeks 2, 5 and 8 with 0.5mg KLH-peptide and Incomplete Freunds Adjuvant. KLH-peptide sera was tested from day 0 and week 10.

Statistical analyses

We performed statistical analyses using a one-tailed Students t-test to obtain p-values on the significance of relative avidity levels below specified thresholds.

2.4 Results

2.4.1 Computational design of epitope-scaffolds

We designed a total of 103 4E10 epitope-scaffolds by the side-chain grafting method (models representative of the variety of scaffold topologies are shown in Fig 2.1).

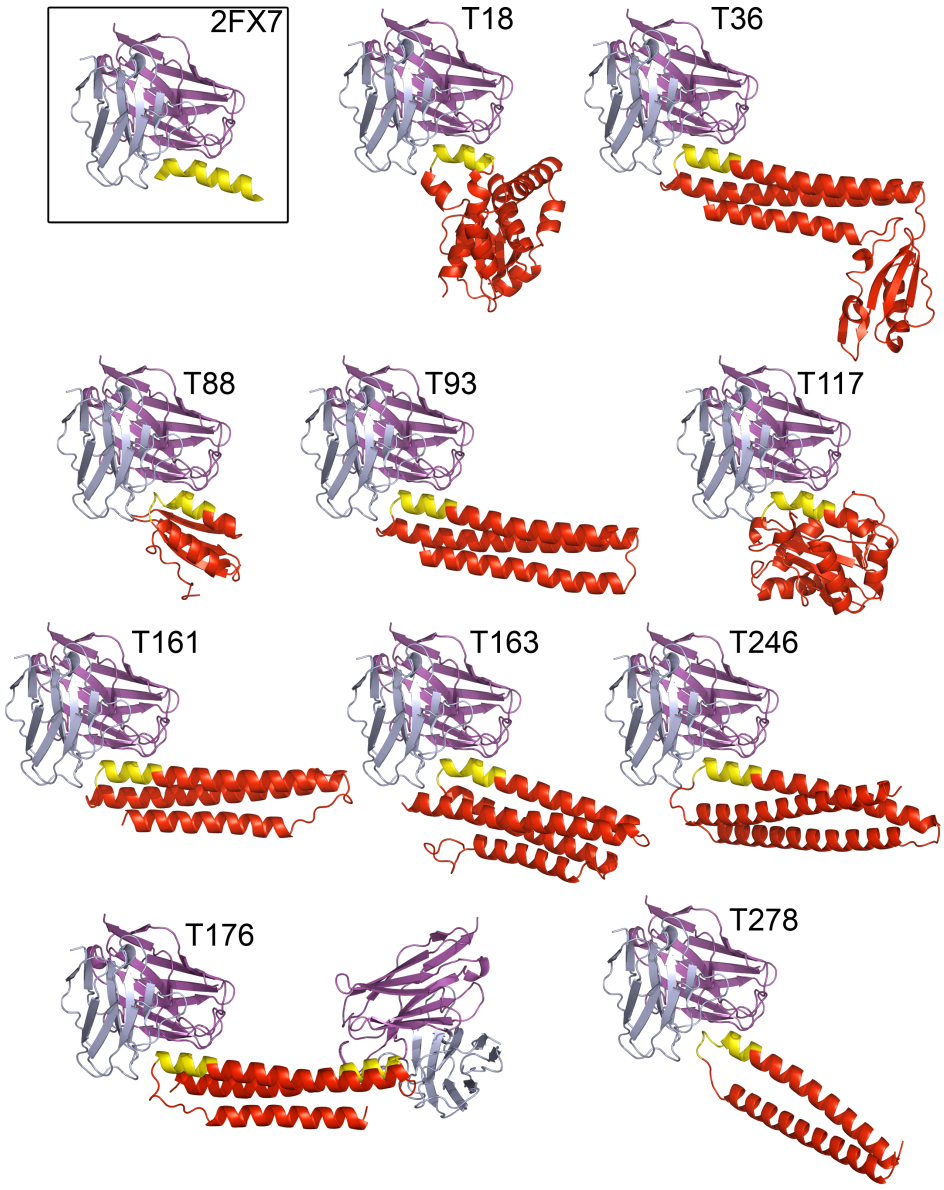


Figure 2.1: Models of 4E10 Fv/epitope-scaffold complexes. The helical 4E10 epitope (yellow) was transplanted from the 4E10 Fab/gp41 peptide complex (boxed at top left) onto different scaffold proteins (red).

This method entailed 'matching' and 'design' stages. In the automated 'matching' stage, a subset of the Protein Data Bank [4] was exhaustively searched for scaffold proteins with a surface-exposed backbone segment similar in conformation to the 4E10 epitope (NWFDITxxLW in PDB ID: 2FX7[61]). The resulting 'hits' were filtered first by backbone clash with the antibody and second by antibody-epitope binding energy (epitope side-chains were transferred to the scaffold and energy minimization [78] was used to optimize antibody and epitope side-chain conformations as well as the antibody-scaffold rigid-body orientation). Scaffolds selected in the 'matching' stage were taken to the semi-automated 'design' stage in which amino acids were designed for scaffold positions outside the epitope. The 'design' stage aimed to ensure scaffold stability and solubility, avoid additional interactions between scaffold and antibody, support the optimized epitope side-chain conformations, eliminate undesired functional sites, and trim extraneous domains.

2.4.2 'Validated' epitope-scaffolds are stable and bind MAb 4E10 with high affinity

Epitope-scaffolds were expressed in *E. coli*, purified by immobilized metal affinity chromatography and SEC, and screened for antibody binding in qualitative SEC assays. Constructs were validated if they expressed as soluble protein (not within inclusion bodies) and were purifiable to stable, soluble species that bound antibody by SEC; non-validated constructs were eliminated from further consideration. A total of 135 constructs were tested for expression (103 designs plus 32 purification tag variants), and of those, 24 different designs (23% of 103 designs) were carried forward for quantitative analysis of bNAb-binding by SPR. Biophysical data are collected in Table 2.3 for the representative set of 10 validated 4E10 epitope-scaffolds depicted in Figure 2.1, and background information on each scaffold is provided in Table 2.4.

Table 2.3: Solution properties of 4E10 epitope-scaffolds.

Epitope-scaffold	Solution Multimer State	Tm (°C)	Fv:epitope-scaffold Binding Stoichiometry	Interaction Parameters (SPR)					
				Analyte: 4E10 Fv Ligand: epitope-scaffold[1]			Analyte: epitope-scaffold Ligand: 4E10 IgG[2]		
				k_{OFF}/k_{ON} (nM)	k_{ON} ($M^{-1}s^{-1}$)	k_{OFF} (s^{-1})	k_{OFF}/k_{ON} (nM)	k_{ON} ($M^{-1}s^{-1}$)	k_{OFF} (s^{-1})
T18	dimer	≥100	2:2	Low Specific Activity[3]			2.75	1.47×10^5	4.05×10^{-4}
T36	monomer	44	1:1	0.014	5.89×10^6	8.34×10^{-5}	0.044	1.95×10^6	8.59×10^{-5}
T88	tetramer	64	2:2	5.6	1:4.0×10^5 2:2.24×10^{-2}	1:4.8×10^{-2} 2:1.1×10^{-3}	0.047	3.94×10^6	1.87×10^{-4}
T93	monomer	79	1:1	0.0075	6.98×10^6	5.23×10^{-5}	0.028	3.17×10^6	8.73×10^{-5}
T117	dimer	56	1:1	Low Specific Activity[3]			≤ 0.01 [4]	--	--
T161	monomer	65	1:1	0.015	7.92×10^6	1.15×10^{-4}	0.11	3.10×10^6	3.43×10^{-4}
T163	monomer dimer	72	multiple species	0.263	2.57×10^6	6.75×10^{-4}		ND	
T176	monomer dimer	48	multiple species	0.14	4.49×10^6	6.18×10^{-4}	0.051	6.44×10^6	3.26×10^{-4}
T246	monomer dimer trimer	62	1:1	3.3	1:2.34×10^6 2:3.06×10^{-3}	1:1.66×10^{-2} 2:2.62×10^{-3}		ND	
T278	monomer	51	1:1	0.89	1:3.13×10^6 2:1.72×10^{-2}	1:3.2×10^{-2} 2:1.63×10^{-3}	1.85	3.71×10^5	6.84×10^{-4}

1: Epitope-scaffold was immobilized by direct amine-coupling on CM5 chip or by biotinylation and capture on SA chip.

2: 4E10 IgG was captured by amine-coupled anti-human IgG.

3: Coupling epitope-scaffold to chip results in an inactive surface, likely due to lysine residues neighboring the epitope.

4: Stated value is an upper limit on the K_D; off-rate was too slow to fit kinetics.

Italicized values: SPR data were fit using a 2-state interaction model; the second on-rate has units of sec^{-1}.

ND: experiment not performed

For all reported values, standard error is $\leq \pm 3$ of the last significant figure.

Table 2.4: Design information on the 4e10 epitope-scaffolds. N Residues reports the size of the parent scaffold protein. N Residues Deleted tells the number of residues deleted from the parent. N mutations counts the mutations to accommodate the 4e10 binding site, remove extra contacts with the 4E10 antibody and enhance protein solubility. Epitope Transplanted reports HIV sidechains on the epitope-scaffold, the dotted positions represent scaffold positions at which HIV sidechains were not transplanted. Epitope RMSD Model reports the root mean square deviation between the backbone of the 4E10 gp41 peptide (PDB ID: 2FX7 [61]) and the parent scaffold. Expected Multimer State reports the expectations based on our analysis of multimer state in the parent scaffold crystal packing, our analysis of the literature regarding the parent scaffold, and on predicted effects on multimerization due to epitope transplantation (e.g. T117 parent was a dimer but we introduced the epitope at the dimer interface so a monomer was expected in the epitope-scaffold). Expected Fv-Binding Stoichiometry reports our expectations based on the number of epitopes transplanted to each monomer (e. g. we transplanted 2 epitopes onto T176) and on the exposure of epitopes on any expected multimers (e. g. T18 is a dimer with both epitopes exposed).

Target	**PDB code/Chain**	**Protein Name**	**Organism**	**N Residues**	**N Residues Deleted**	**N Mutations**	**Epitope Transplanted**	**Epitope RMSD Model (Å)**	**Expected Multimer State & Fv-Binding Stoichiometry**
T18	1Z6N/A	Hypothetical protein PA1234	*P. aeruginosa*	167	-	25	`GWFDIT..LW`	0.32	Dimer, 2:2
T36	1ISE/A	Ribosome recycling factor	*E. coli*	185	-	11	`NWFDIS..LW`	0.85	Monomer, 1:1
T88	1VI7/A	Hypothetical protein yigZ	*E. coli*	71	137	11	`NWFDIT..LW`	0.39	Monomer, 1:1
T93	1ISE/A	Ribosome recycling factor	*E. coli*	116	69	11	`NWFDIS..LW`	0.85	Monomer, 1:1
T117	1XIZ/A	Putative phosphotransferase	*S. typhimuruim*	159	-	10	`NWFDIT..LW`	0.32	Monomer, 1:1
T161	1IS1/A	Ribosome recycling factor	*V. parahaemolyticus*	114	72	6	`NWFDIS..LW`	0.86	Monomer, 1:1
T163	1K05/B	Focal Adhesion kinase 1	*H. sapiens*	162	7	22	`.WFTIT..LW`	0.33	Monomer, 1:1
T176	1ISE/A	Ribosome recycling factor	*E. coli*	114	69	12	`NWFDIS..LW`	0.84	Monomer, 1:2
T246	1EZ3/A	Syntaxin-1A	*R. norvegicus*	127	-	37	`NWFDIS..LW`	1.1	Monomer, 1:1
T278	1TJL/C	DnaK suppressor protein	*E. coli*	79	69	6[a]	`NWFDIT..LW`	0.25	Monomer, 1:1

To facilitate interpretation of bNAb-binding experiments, the epitope-scaffold solution multimer state and stoichiometry of Fv-binding were assessed by static light scattering (SLS) and SEC. Examples of SEC data are given in Figure 2.2. Solution multimer state and Fv-binding stoichiometry accorded with expectation (monomer and 1:1, or dimer and 2:2 for T18) for 5 of 12 epitope-scaffolds (T18, T36, T93, T161, T278). The other epitope-scaffolds formed unexpected multimers that dissociated in the presence of Fv, suggesting that multimers self-associating through the hydrophobic 4E10 epitope were in equilibrium with rare monomers. Aggregation was noted previously for gp120 with engrafted MPER sequences [89]. Multimerization of some 4E10 epitope-scaffolds will be addressed in detail in a separate manuscript (Xu et al., in preparation).

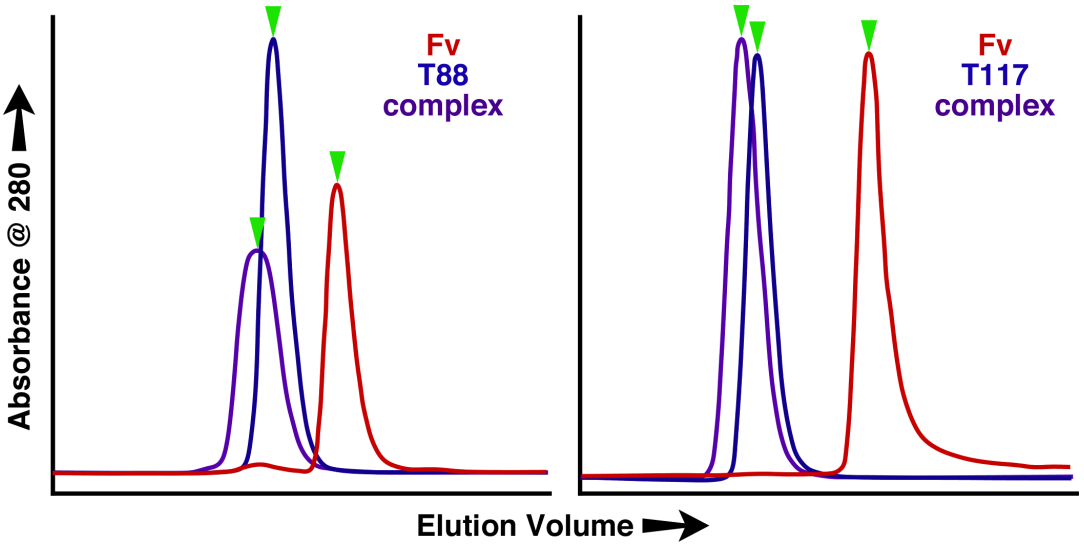


Figure 2.2: Size exclusion chromatography analysis of epitope-scaffolds in complex with 4E10. Epitopes-scaffolds (blue), 4E10 Fv (red) and complexes (purple) are shown for T88 (left) and T117 (right). Molecular weights (Mr) for the species at absorption peaks (arrows) are T88: ∼40kD (SEC) or 32.2kD (SLS); T88/Fv complex: ∼54kD (SEC) or 67.2kD in static light scattering (SLS); T117: ∼35kD (SEC) or 35.4kD (SLS); T117/Fv complex: ∼45kD (SEC) or 45.2kD (SLS); and Fv: ∼17kD (SEC) or 26.2kD (SLS). The predicted Mr for a T88 monomer is 8.4kD, demonstrating that T88 is tetrameric in solution and binds Fv to form 2:2 complexes. The predicted Mr for a T117 monomer is 18.4kD, demonstrating that T117 is dimeric in solution and binds Fv to form 1:1 complexes. Deviations between estimates reflect the greater sensitivity of SEC to departures from sphericity.

Epitope-scaffold affinities for MAb 4E10 were generally quite high. Antibody binding affinity and kinetics were assessed by SPR in two formats: with 4E10 Fv in solution as 'analyte' and epitope-scaffolds coupled to the sensor chip as 'ligand' (to assess interactions without avidity), and with 4E10 IgG captured on the chip and epitope-scaffold in solution (to allow avidity). Examples of SPR data and fits are shown in Figure 2.3, and conditions for each SPR experiment are given in Table 2.6 and 2.7. With epitope-scaffolds on the chip, equilibrium dissociation constants (K_D) ranged from 8 pM to 6 nM, and all epitope-scaffolds bound 4E10 Fv more tightly than 4E10 peptides (K_D = 10-20 nM [90, 64]), gp41 (K_D = 32-35 nM [91]) or gp140 trimers from isolate SF162 (K_D = 130 nM [64]). The range of K_Ds was similar with epitope-scaffolds as analytes (22 pM to 3 nM). Monomeric epitope-scaffolds exhibited similar K_Ds in both formats, while 120-fold tighter binding due to bivalent avidity was detected for T88 (K_D = 47 pM vs 5.6 nM) and 2.7-fold tighter binding was detected for T176 (K_D = 51 pM vs 140 pM), consistent with stoichiometry data. In both formats, variations in both k_{on} and k_{off} contributed to variations in K_D.

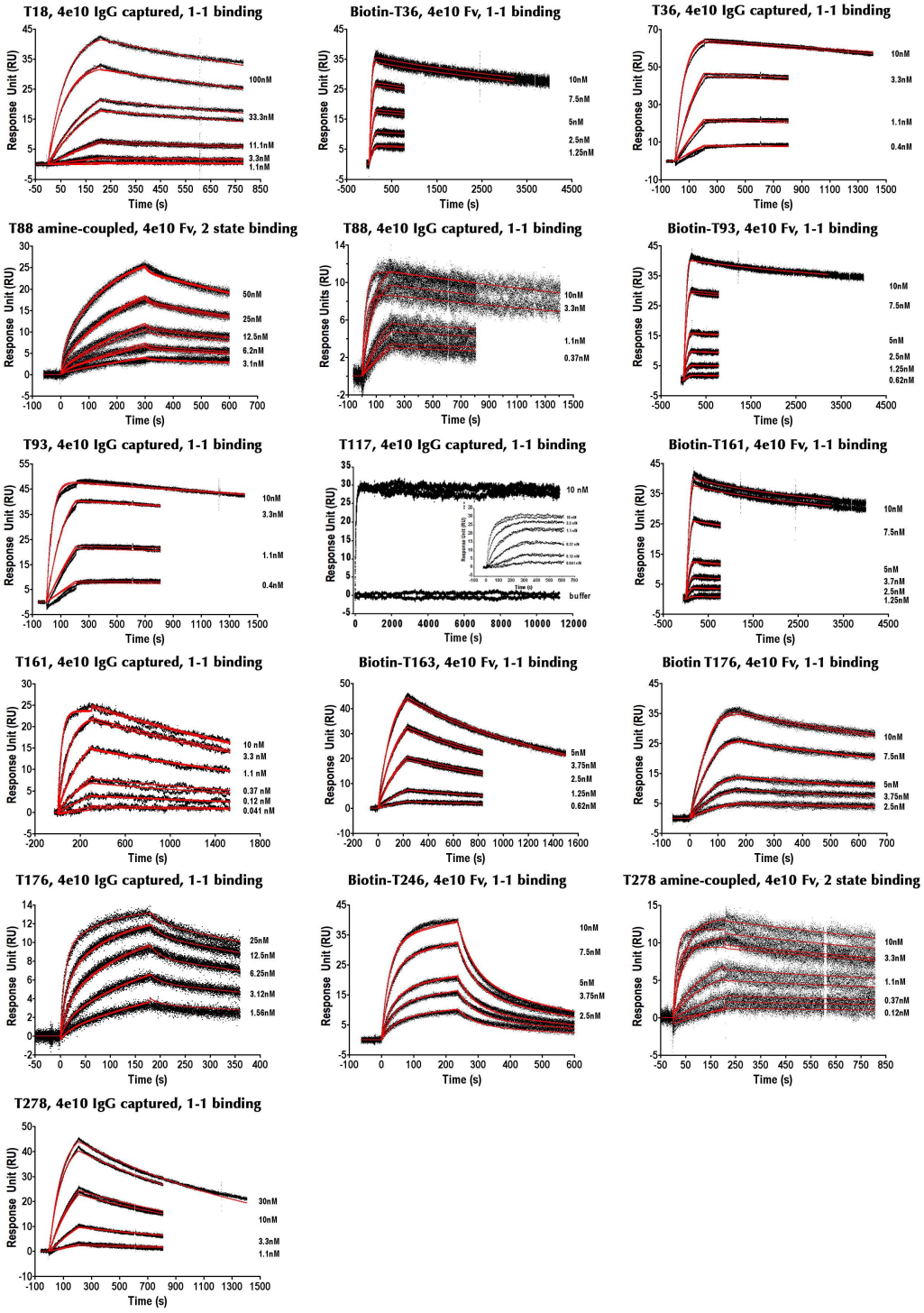


Figure 2.3: SPR binding kinetics of 4E10 Fv/IgG with epitope-scaffolds or biotinylated versions. Double reference-subtracted SPR responses in black with the corresponding kinetic fits in red.

Epitope-scaffolds showed a range of thermal stabilities assessed by circular dichroism temperature melt analysis (T_m = 44 ^{o}C to $T_m \leq$ 100 ^{o}C), though a majority (60%) were very stable with Tm > 60 ^{o}C, and a large majority (80%) were stable with Tm > 50 ^{o}C.

2.4.3 Crystallographic analyses reveal high epitope structural mimicry by 4E10 epitope-scaffolds

To assess epitope structural mimicry, crystal structures were determined for six different epitope-scaffolds, four of unliganded epitope-scaffolds, and two of complexes with a 4E10 Fv construct biochemically validated elsewhere [64] (Figure 2.4). Crystallization conditions are in Table 2.2; X-ray diffraction and refinement statistics are in Table 2.4; stereo images of electron density maps are given in Figure 2.11.

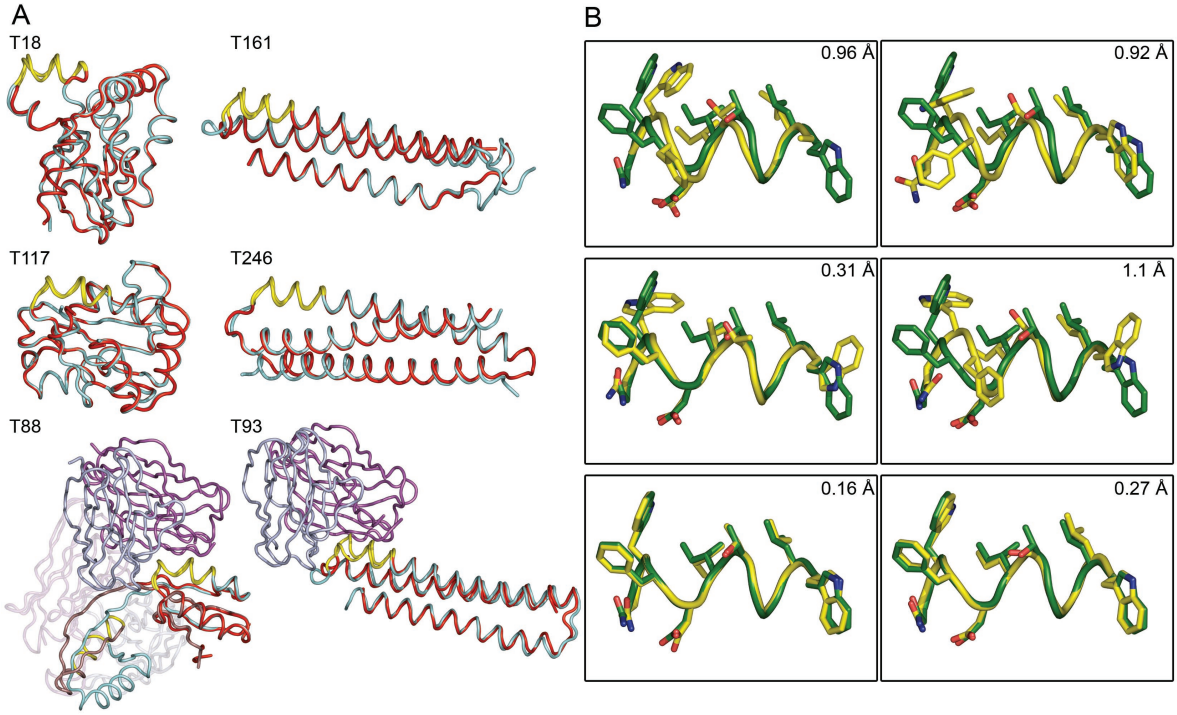


Figure 2.4: Crystallographic analysis of 4E10 epitope-scaffolds.(A) Crystal structures of six epitope-scaffolds (cyan tubes) superimposed with design models (red tubes), with the epitope region colored yellow. Four structures are unliganded and two are complexed with 4E10 Fv W(H100A)A (H-chain, light-blue; L-chain, magenta). Resolution of structures was 2.3 (T18, PDB ID: 3LEF), 2.0 (T161, PDB ID: 3LF9), 1.9 (T117, PDB ID: 3LF6), 2.5 (T246, PDB ID: 3LG7), 2.65 (T88-Fv, PDB ID: 3LH2), and 2.8 (T93-Fv, PDB ID: 3LHP) angstroms. (B) Epitope structure (yellow tubes and sticks) from each of the epitope-scaffold crystal structures, superimposed with the 4E10 Fab-bound structure of a gp41 peptide (green tube and sticks) from PDB ID: 2FX7[61]. Backbone RMSD values between epitope-scaffold and peptide are indicated for each superposition.

The overall conformations of the epitope-scaffolds agreed well with design models in five of six cases (Figure 2a), with backbone RMSD values between models and structures of 0.47, 0.50, 0.79, 1.14, and 1.26 Å for epitope-scaffolds T117, T18, T246, T93, and T161, respectively. The structure of the T88-Fv complex revealed an unintended domain-swapped dimer binding two Fv molecules, consistent with the 2:2 stoichiometry measured by SEC and SLS. Nevertheless both T88-Fv complexes were very similar to the design model; ignoring two residues that bridge the T88 dimer, backbone RMSD between the two Fv-monomer complexes and the model were 0.6 Å and 0.5 Å respectively.

Most importantly, epitope mimicry was excellent in both Fv-bound and unbound epitope-scaffolds (Figure 2.4B). Backbone RMSD between epitope-scaffold and the 4E10-bound conformation of the gp41 peptide [61], computed over epitope residues 672-680, was 0.99, 0.31, 0.99, and 1.1 Å for the unbound structures of T18, T117, T161, and T246 respectively, 0.16 Å for each monomer in the T88-Fv complex, and 0.31 Å in the T93-Fv complex. B-factor analysis indicated that the epitope backbones were generally as well-ordered as the epitope-scaffolds overall (Figure 2.5).

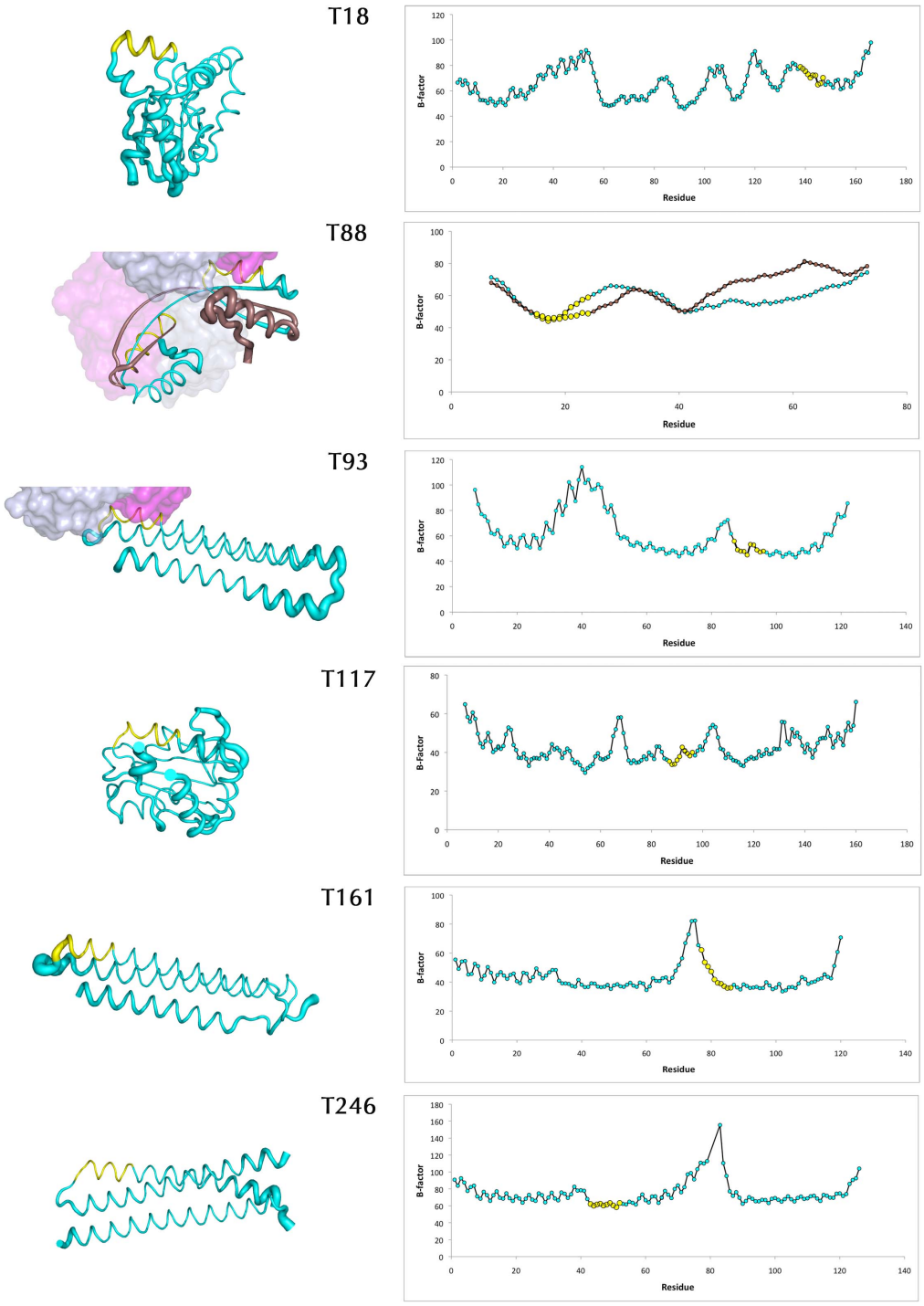


Figure 2.5: B-factor analysis of epitope-scaffolds. Crystal structures of the epitope scaffolds are shown in cyan tubes with the epitope region in yellow; in epitope-scaffold complexes the Fv H and L chains are shown as grey and magenta surfaces, respectively. The width of the tube scales with the crystallographic B-factor (C-alpha atoms). Plots of C-alpha B-factors versus residue number are shown in the right. In each case, data shown are representative for all structures within the asymmetric unit.

Side-chain mimicry was reduced in the unliganded structures (Figure 2.4B), but high affinities for Fv indicated that appropriate side-chain conformers were attainable. Indeed, inter-molecular contacts within the unliganded crystals stabilized different epitope side-chain conformations at some positions (Figure 2.12). The T88-Fv and T93-Fv complexes demonstrated precise mimicry at the side-chain level, with all-atom epitope RMSDs of 0.48 and 0.36 Å respectively; these complexes reproduced nearly all atomic contacts across the bNAb-peptide interface [61, 62, 64] but made some additional contacts (Tables 2.5, 2.8, 2.9 and Figure 2.6).

Table 2.5: Buried surface areas of the epitope-scaffolds and the combined site of 4E10 Ab. The buried surface areas were computed with the NACCESS [92] using a standard 1.4 Å radii water probe.

Complex	Peptide/epitope-scaffold surface area occluded by 4e10(Å)	4e10 Surface Area occluded by peptide/epitope-scaffold(Å)	Epitope-scaffold additional surface area relative to the 2FX7 peptide(%)	4e10 additional surface area relative to the 2FX7 structure(%)
2FX7	817	649	-	
T88-Fv	980	869	19.9	33.8
T93-Fv	931	822	13.9	26.6

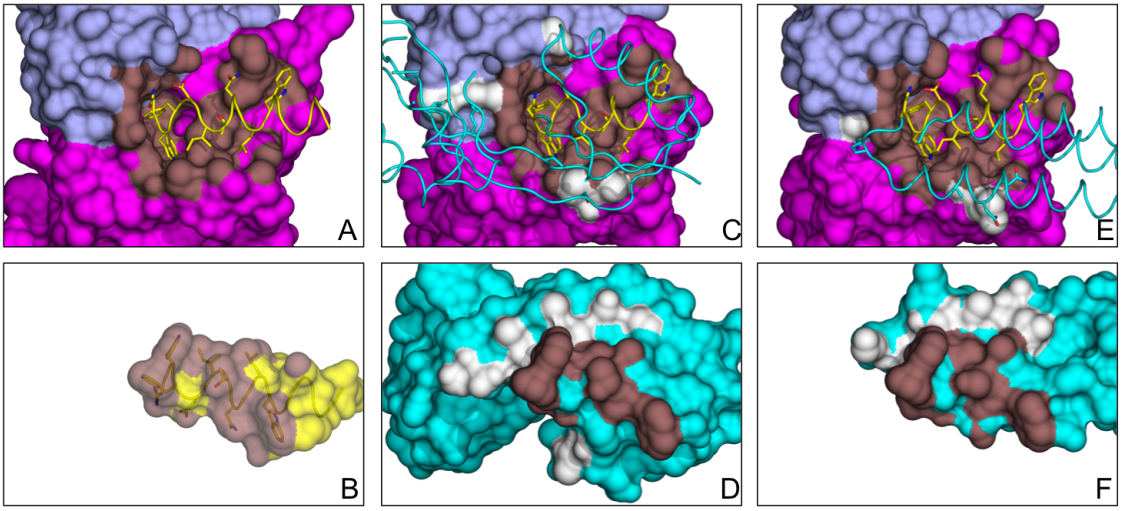


Figure 2.6: Analysis of contact surfaces in the 4E10 Fab/gp41[61] and the 4e10 Fv/epitope-scaffold complexes. Panels A-B show the 4E10 Fab/gp41 interface, looking through the gp41 (tube/stick) toward the Fab surface (A) or toward the peptide surface from the Fab (B). Panels C-D show the 4E10 Fv/T88 interface, looking through T88 toward the Fv (C) or toward T88 from the Fv (D). Panels E-F show the 4E10 Fv/T93 interface, with views analogous to those in C-D. The epitope segments of gp41, T88 and T93 are shown in yellow tube (backbone) and stick (side chains), otherwise the T88 and T93 backbones are shown in cyan tube. Atoms on the 4E10 epitope that contact the Fab or Fv are colored brown (B, D, F), as are the corresponding contact atoms on the Fab or Fv (A, C, E). Atoms on the Fv or epitope-scaffold involved in non-epitope contacts are colored white (C-F), except that the epitope-scaffold side chains involved in non-epitope contacts are shown in cyan sticks in C and E. 4E10 Fab or Fv H-chain is magenta, L-chain is grey. Most of the additional contacts for T88 involved the domain-swapping 'bridge' between the two monomers, and the additional contacts for T93 were due to small conformational changes in the epitope-scaffold and Fv compared to the design model. Such additional contacts might enhance binding affinity for bNAb 4E10. Graphics were prepared with Pymol [93].

2.4.4 Epitope-scaffolds as reagents reveal that neutralizing activity in HIV+ sera has 4E10-like helical specificity while non-neutralizing anti-peptide sera does not

Sather et al.[66] recently found evidence for the presence of 4E10-like antibodies in the serum of one HIV+ individual ("VC10028"). Support for this finding came primarily from the fact that a 4E10 peptide, but not a scrambled 4E10 peptide, could inhibit a fraction of the neutralizing activity of this serum in an *in vitro* neutralization assay. There are many potential antibody epitopes on the 4E10 peptide, especially given that the peptide likely samples many different conformations in solution, so the precise structural epitope(s) of the neutralizing antibodies in serum VC10028 could not be identified. Having demonstrated epitope structural mimicry crystallographically (Figure 2.4), we employed 4E10 epitope-scaffolds as reagents to determine whether the epitope conformation stabilized on the scaffolds was a target of the cross-neutralizing antibodies in VC10028. Several epitope-scaffolds inhibited the neutralizing activity of the sera (Figure 2.7A, B), but no inhibition was detected by a 'dead-epitope' version of one scaffold (X36) in which the 'WF' residues of the epitope were mutated to 'ED' (Figure 2.7B). These results demonstrated that neutralizing antibodies targeting the helical epitope structure targeted by MAb 4E10 were present in the VC10028 sera.

In contrast, non-neutralizing rabbit sera elicited by a 4E10 peptide-KLH conjugate had no significant affinity for any of several epitope-scaffolds tested, even though the peptide-KLH sera bound avidly to 4E10 peptide itself (Figure 2.8A). This result demonstrated that while the 4E10 peptide-KLH conjugate was immunogenic (one of three rabbits produced peptide-specific antibodies), it did not elicit antibodies with structural specificity similar to 4E10. Indeed, comparison of peptide-KLH sera and MAb 4E10 binding to ALA-scanned peptides (Figure 2.8B,C) confirms that the specificities are very different; MAb 4E10 binding is greatly diminished by mutants W672A, F673A, T676A, and L679A that remove critical contacts in the MAb/peptide crystal structure [61, 62, 64], but peptide-KLH sera binding to those mutants is undiminished. Together, these results showed how epitope-scaffolds may be useful as reagents to identify the presence or absence of precise structural specificities in sera, and implied that epitope-scaffolds will be useful as bait to isolate antibodies targeting

pre-defined structural epitopes.

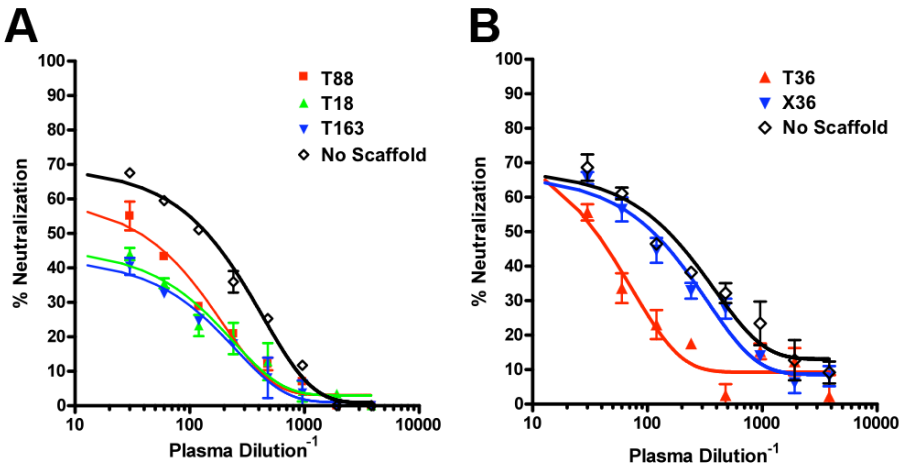


Figure 2.7: 4E10 epitope-scaffolds inhibit neutralizing activity in HIV+ serum VC10028[66].(A and B) Neutralization of JR-FL pseudovirus by serial dilutions of serum VC10028 in the absence (open diamonds) or presence (solid shapes) of different 4E10 epitope-scaffolds at a concentration of 10 μg/mL epitope-scaffolds. All epitope-scaffolds partially inhibit the neutralizing activity except a 'dead-epitope' epitope-scaffold (X36, blue inverted triangles in (B)). Data are representative of 3 separate experiments; error bars represent means $\pm$ s.d.

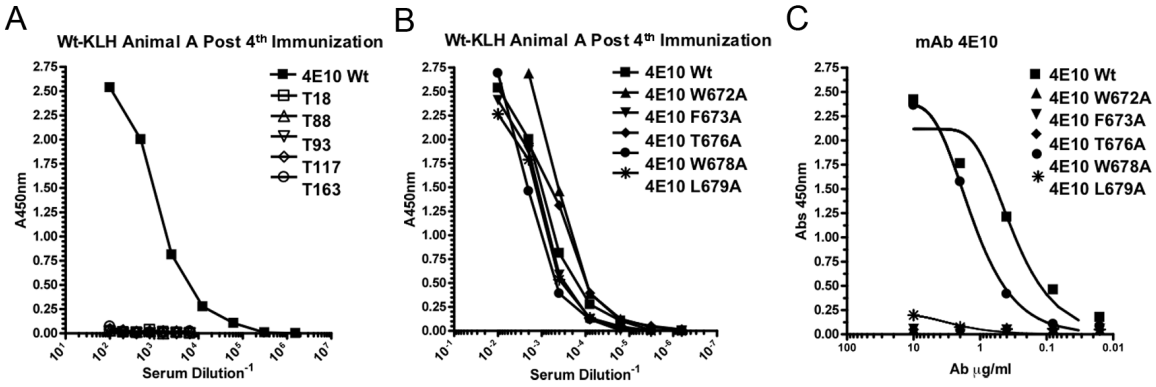


Figure 2.8: Structural specificities of bNAb 4E10 and anti-4E10-peptide rabbit serum are dramatically different. (A) ELISA of anti-peptide serum binding to 4E10 peptide and different epitope-scaffolds. Data representative of 2 experiments; error bars represent means ± s.d. (B) ELISA of anti-peptide serum binding to alanine-mutant 4E10 peptides. Data representative of 3 experiments; error bars represent means ± s.d. (C) ELISA of bNAb 4E10 binding to alanine-mutant 4E10 peptides. Data representative of 3 experiments; error bars represent means ± s.d.

2.4.5 An epitope-scaffold elicited a structurally specific anti-epitope response

The fact that the 4E10-like antibodies in serum VC10028 had structural specificity similar to MAb 4E10, while non-neutralizing antibodies in peptide-KLH sera did not, suggested that neutralization via the 4E10 epitope may require antibodies with 4E10-like structural specificity, and underscored the importance of testing epitope-scaffolds for their ability to re-elicit such structural specificity. To this end we carried out immunization experiments with specific endotoxin-free preparations of one 4E10 epitope-scaffold (T88). Two groups of 3 rabbits were immunized with T88, one group with QS21 as adjuvant, and the other with PEI. One animal from each group developed antibodies that bound to HXB2 gp41, and this gp41-reactivity was inhibited by wild-type 4E10 peptide, demonstrating specific reactivity to the 4E10 epitope sequence (Figure 2.9). Further, the gp41-reactivity of the anti-T88 sera from these two rabbits was inhibited by three different epitope-scaffolds tested (T36, T161, and T163) but not by a 'dead-epitope' version of the T36 epitope-scaffold (Figure 2.9), demonstrating specific reactivity to the helical structure of the 4E10 epitope stabilized on the scaffolds.

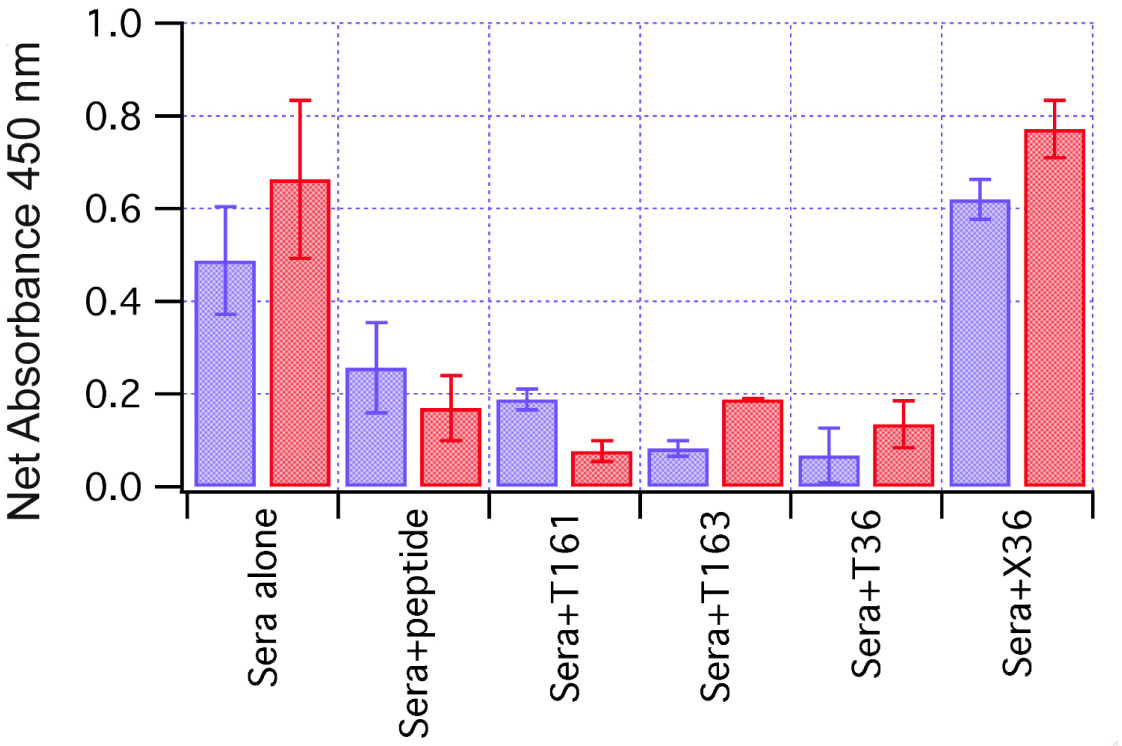


Figure 2.9: ELISA binding of anti-T88 rabbit sera to gp41 in presence of peptide and epitope-scaffold competitors. Anti-T88 sera from two rabbits (blue, PEI adjuvant, five immunizations; red, QS21 adjuvant, six immunizations) binds to gp41 (”Sera alone”). Addition of peptide (”Sera+peptide”) or different scaffolds (”Sera+T161”, ”Sera+T163”, ”Sera+T36”) significantly reduces gp41-binding, but addition of a ’dead-epitope’ epitope-scaffold (”Sera+X36”) does not. ELISA response of pre-bleed rabbit sera to gp41 was subtracted from the data. Data were averaged from 2 or 4 experiments; error bars represent means ± s.d.

To assess the fine specificities of the anti-T88 sera and MAb 4E10 across the helical epitope, we measured the binding of both sera and MAb to alanine-substituted 4E10 peptides using a multiplexed Luminex assay [94]. The results are displayed in Figure 2.10 as 'relative avidity' in which the avidity for each ala-mutant is divided by the avidity of MAb 4E10 for the wild-type peptide. A relative avidity significantly less than 1.0 indicates engagement of a particular side-chain, as mutations to alanine should not detract from peptide helical propensity. Strikingly, the results indicated that both T88 sera strongly engaged most epitope side-chains transplanted onto the T88 scaffold, as relative avidities for ala-mutants at gp41 positions 672-675 (WFDI) and L679 were less than 0.3 ($p < 0.03$ in all cases except sera2 at I675 where relative avidity was less than 0.4 with $p = 0.02$). Two transplanted epitope side-chains (T676 and W680) were engaged by one of the sera but not both; T676 relative avidity was 0.33 ± 0.28 for sera1 (engaged with relative avidity < 0.7, $p = 0.02$) but 1.61 ± 0.33 for sera2, and W680 relative avidity was 0.49 ± 0.21 for sera2 (engaged with relative avidity < 0.8, $p = 0.01$) but 0.74 ± 0.50 for sera1. Relative avidities for MAb 4E10 showed a similar overall pattern of side-chain engagement, except that MAb 4E10 did not engage I675 (relative avidity, 1.35 ± 0.18) and did strongly engage T676 (relative avidity, 0.08 ± 0.05). The data for MAb 4E10 agreed with previous competitive ELISA experiments [90] except at positions 675 and 680 where Luminex indicated no MAb engagement but competitive ELISA [90] indicated moderate engagement. As expected, neither of the two T88 sera nor MAb 4E10 had significantly altered avidity for ala-mutations of gp41 residues N677 and W678 that are located on the opposite face of the helix from the WFDITxxLW epitope of 4E10 and were not transplanted onto the T88 scaffold.

Gp41-specific IgG was purified from the sera and tested for fine specificity and neutralization. The fine specificity of the gp41-specific IgG was consistent with that of the sera in Figure **??**. Neither the sera nor purified and concentrated anti-gp41 IgG reproducibly neutralized HIV-1 in an *in vitro* neutralization assay (not shown).

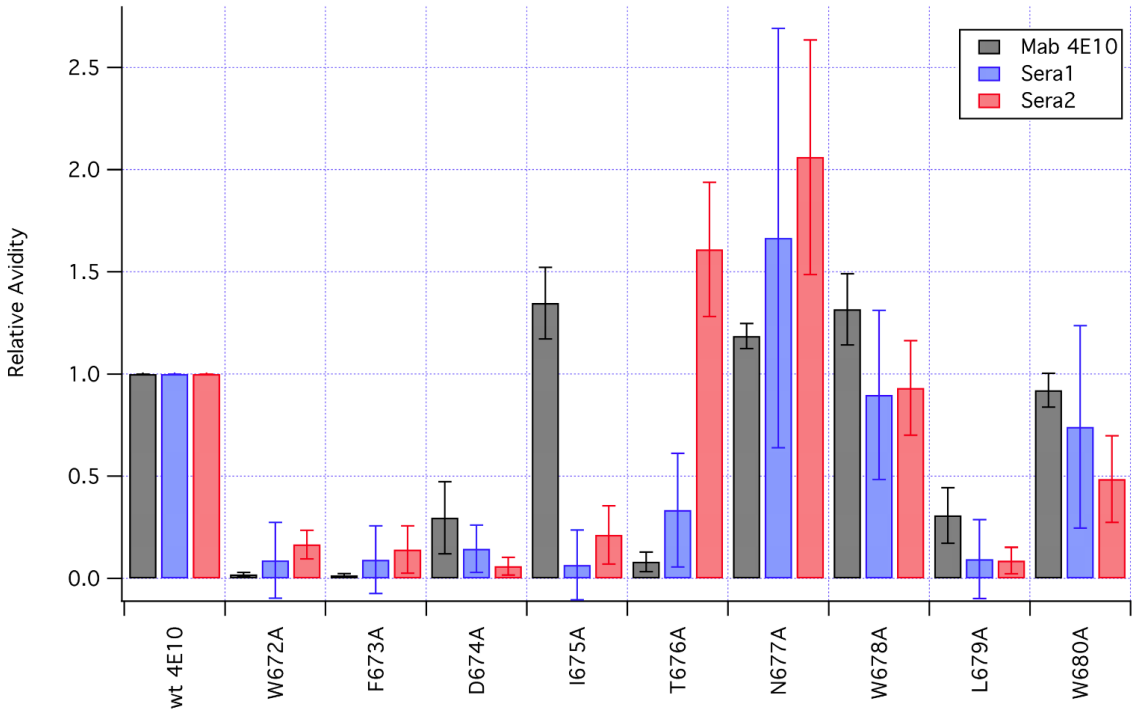


Figure 2.10: Luminex analysis of bNAb 4E10 and anti-T88 rabbit sera binding to ala-scanned gp41 peptides on beads.Relative avidity for bNAb 4E10 (black) and two different anti-T88 rabbit sera (blue, PEI adjuvant; red, QS21 adjuvant) quantifies the avidity of each antibody-peptide or antisera-peptide pair relative to bNAb 4E10 binding to wild-type 4E10 peptide. Both anti-T88 antisera exhibit low relative avidity for mutants W672A, F673A, D674A, I675A, and L679A, indicating that the antisera strongly engaged the 4E10 side-chains at those positions on the helical epitope. Data were averaged from 5 experiments each run in duplicate; error bars represent mean ± s.d.

2.5 Discussion

Here, we presented a new technique, computational design of epitope-scaffolds, to design immunogens that preserve the structure of a conserved epitope and expose that structure on a variety of different protein scaffolds with different immunologic backgrounds. We transplanted a conserved HIV epitope to multiple non-HIV proteins, and succeeded to produce small, stable, soluble proteins with high affinity for the target monoclonal (K_Ds < 1 nM). Indeed, several epitope-scaffolds bound MAb 4E10 up to 1000-fold more strongly than cognate peptide ($K_D \sim 10$ pM for epitope-scaffolds vs 10 nM for peptide). Crystal structures of multiple scaffolds demonstrated near perfect mimicry of the desired epitope backbone conformation (RMSDs ≤ 1 Å). We demonstrated the utility of epitope-scaffolds as serologic reagents to identify the presence or absence of antibodies targeting specific structural epitopes; this diagnostic utility conceptually extends to the use of epitope-scaffolds as bait for isolating such antibodies from libraries. Most importantly, we demonstrated that antibodies specific for the helical structure and the conserved side-chains of the 4E10 epitope could be elicited by immunization with one such 4E10 epitope-scaffold. In contrast, immunization with peptide-KLH failed to elicit this desired specificity, and previous studies showed that several different constructs competent for 4E10 binding failed to elicit any detectable reactivity to the 4E10 linear sequence [67, 68, 89]. This example suggests the potential power of epitope-scaffolding; with the technique described here, many types of contiguous epitopes could be transplanted onto scaffolds for structural stabilization and immune presentation, and epitope-structure-specific immune responses might be further enhanced by glycan- or PEG-masking of the scaffold, by multimeric, oriented display of epitope-scaffolds on particles, by heterologous prime-boost immunizations, or combinations thereof. We also note that the epitope-scaffold concept is not limited to scaffolding of antibody epitopes, it could also prove useful to study, or design inhibitors for, peptide-protein and protein-protein interactions.

To our knowledge, the antibodies elicited by T88 are the first induced by vaccination to possess structural specificity highly similar to that of 4E10. This is an important advance for HIV vaccine design, because the helical specificity possessed by 4E10 and also

the neutralizing sera in VC10028 [66] is likely necessary for neutralization via the 4E10 epitope. The T88 antisera elicited by these first-generation designs did not neutralize HIV, however, indicating that helical specificity and the engagement of the conserved side-chains of the 4E10 peptide epitope may not be sufficient. This result is consistent with recent studies demonstrating that specific hydrophobic residues on the tip of the 4E10 CDRH3 are required for maximum potency neutralization: alanine [95, 63, 64] and aspartic acid [63] substitutions at the tip do not significantly reduce gp41 or epitope-peptide binding, but they do reduce both lipid-binding [95, 63, 64] and the extraction of the epitope from the membrane [64], either mechanism potentially contributing to neutralization potency. As mentioned above, the tip of the 4E10 CDRH3 does not contact the peptide epitope used as the basis of the design of these first-generation epitope-scaffolds [61, 62, 64]; further, no contacts with the CDRH3 tip are seen in Fv 4E10/epitope-scaffold complex structures as this was not a design criteria. Next-generation epitope-scaffolds incorporating higher-order structural complementarity, such as viral membrane lipid structures, will be employed to improve on this approach.

In summary, we introduced a computational methodology that allows the faithful transplantation of conformational epitopes into distinct protein platforms, and demonstrated that this leads to the appropriate stimulation of humoral responses to a conserved, but intrinsically poorly immunogenic viral epitope. Epitope-scaffolds therefore expand the universe of immunogens available for vaccine research. For the case of 4E10, computationally designed scaffolds can next be presented to the immune system in the context of lipid membrane surfaces, in combination with specific adjuvants to elicit antibodies with expanded recognition and neutralization properties.

2.6 Acknowledgements

We thank Della J. Friend for assistance with SPR and for comments on the manuscript. We thank Peter Kwong, Richard Wyatt, Gilad Ofek, and Javier Guenaga for discussions on scaffolding. This work was funded by a grant from the Bill and Melinda Gates Foundation CAVD to L.S., and B.E.C. was supported by a fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/32958/2006).

William R. Schief (W.R.S.), Leonidas Stamatatos (L.S.), Roland K. Strong (R.K.S.), and David Baker (D. B.) conceived and supervised the study. Yih-En Andrew Ban (Y.-E.A.B.), and W.R.S. wrote scaffold design software; Bruno Emanuel Correia (B.E.C), Y.-E.A.B., Chris Carrico and W.R.S. designed scaffolds; Erica Boni (E.B.), Camille Zenobia, Katherine Y. Burke, Tyler Bradley-Hewitt, and Jessica F. Bruhn-Johannsen produced proteins and performed SLS and SEC analysis; Hengyu Xu (H.X.), E.B., and Oleksandr Kalyuzhniy performed SPR analysis; H.X. and E.B. performed CD melt analysis; Margaret A. Holmes (M.A.H.) crystallized many scaffolds and determined six crystal structures; B.E.C and W.R.S. analyzed crystal structures; Katharine Ellingson and Zane Kraft performed ELISA and Luminex analysis; D. Noah Sather performed neutralization competition assays; W.R.S. contributed statistical analysis; B.E.C., R.S., L.S. and W.R.S. wrote the paper; all authors commented on the manuscript.

2.7 Supplemental Material

2.7.1 SPR experimental protocol

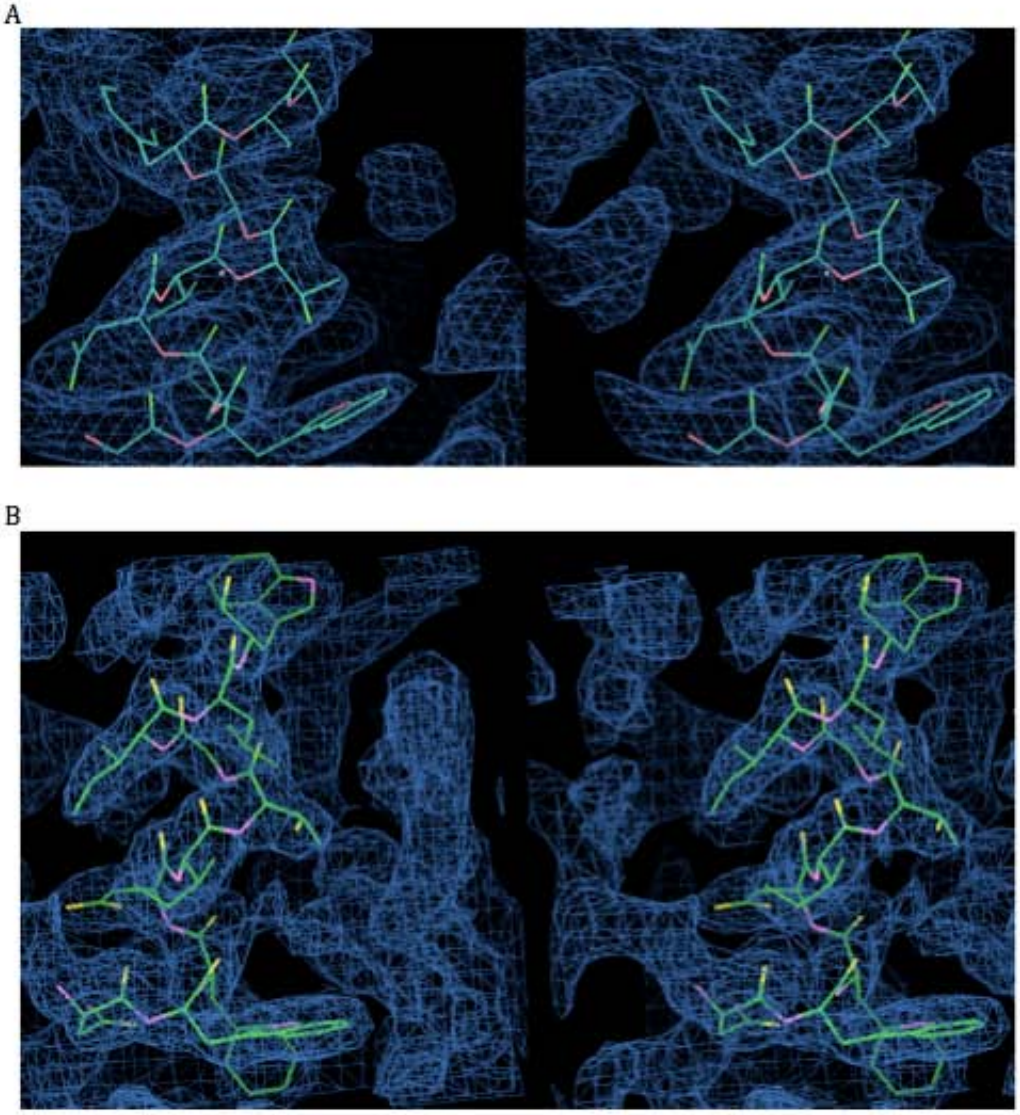
Table 2.6: Surface Plasmon Resonance methods to measure affinities of epitope-scaffolds. 4E10 Fv analyte binding to immobilized epitope-scaffolds.

				Immobilized Epitope Scaffold Ligand				4E10 Fv Analyte				Regeneration			Analysis	
	Biacore Instrument	Chip Type	Running Buffer	Flow Rate (µL/ min)	Concentration (µg/mL)	Injection Time (s)	Immobilized Level (RU)	Flow Rate (µL/ min)	Concentration	Injection Time (s)	Dissociation Time (s)	Flow Rate (µL/ min)	Injection Time (s)	Regeneration Buffer	Model	Software
T18	T100	SA	HBS-EP+	10	≈10	300	500	50	Fv4E10 4.5 µM IgG4E10 500 nM	120	180	50	5	10mM Glycine, pH 1.5	low specific activity	
T36	T100	SA	HBS-EP+	10	≈10	12	44	50	10, 7.5, 5, 2.5, 1.25 nM in duplicate	180	3800 highest concentration, 600 for rest	50	5	10mM Glycine, pH 1.5	1:1	BiaEval 2.0
T88	T100	CM5	HBS-EP+	10	5 µg/mL in NaOAc pH 4.5	420	200	50	50, 25, 12.5, 6.2, 3.1 nM in duplicate	300	300	50	5	10mM Glycine, pH 1.5	2 state	BiaEval 2.0
T93	T100	SA	HBS-EP+	10	≈10	42	60	50	10, 7.5, 5, 2.5, 1.25, 0.62 nM in duplicate	180	3800 highest concentration, 600 for rest	50	10	10mM Glycine, pH 1.5	1:1	BiaEval 2.0
T117	T100	SA	HBS-EP+	10	≈5	60	100	50	Fv4E10 40 nM Fv4E10 200 nM	120	300	50	5	10mM Glycine, pH 1.5	low specific activity slow off rate	
T161	T100	SA	HBS-EP+	10	≈10	42	94	50	10, 7.5, 5, 3.7, 2.5, 1.25 nM in duplicate	180	3800 highest concentration, 600 for rest	50	10	10mM Glycine, pH 1.5	1:1	BiaEval 2.0
T163	T100	SA	HBS-EP+	10	≈5	100	175	50	5, 3.75, 2.5, 1.25, 0.62 nM in duplicate	240	1200 highest concentration, 600 the rest	50	5	10mM Glycine, pH 1.5	1:1	BiaEval 2.0
T176	T100	SA	HBS-EP+	10	≈5	55	60	50	10, 7.5, 5, 3.75, 2.5 nM in duplicate	180	480	50	5	10mM Glycine, pH 1.5	1:1	BiaEval 2.0
T246	T100	SA	HBS-EP+	10	≈5	36	180	50	10, 7.5, 5, 3.75, 2.5 nM in duplicate	240	480	50	5	10mM Glycine, pH 1.5	2 state	BiaEval 2.0
T278	T100	CM5	HBS-EP+	10	10 µg/mL in NaOAc pH 4.5	420	*225*	50	25, 12.5, 6.25, 3.12, 1.56 nM in duplicate	180	180	50	25	0.085% H3PO4	2 state	BiaEval 2.0

Table 2.7: Surface Plasmon Resonance methods to measure affinities of epitope-scaffolds. Epitope-scaffolds analytes binding to captured 4E10 IgG .

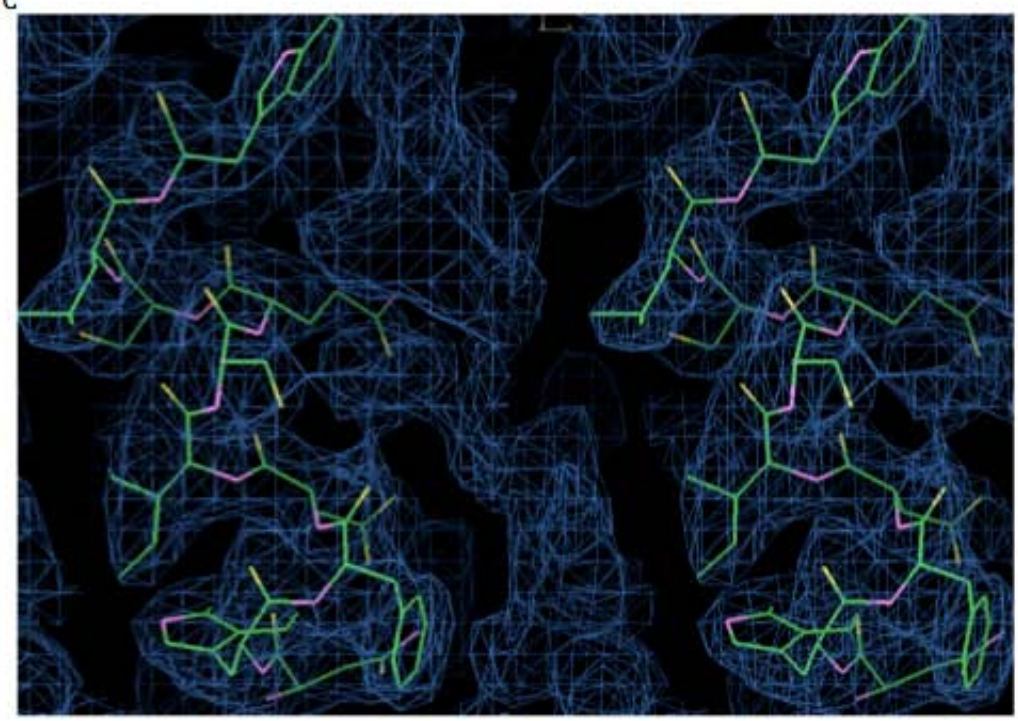
					Captured 4E10 IgG Ligand				Epitope-Scaffold Analyte				Regeneration			Analysis	
	Biacore Instrument	Chip Type	Running Buffer	Immobilized Anti-IgG Level (RU)	Flow Rate (µL/ min)	Concentration (µg/mL)	Injection Time (s)	Captured Level (RU)	Flow Rate (µL/ min)	Concentration	Injection Time (s)	Dissociation Time (s)	Flow Rate (µL/ min)	Injection Time (s)	Regeneration Buffer	Model	Software
T18	T100	CM5	HBS-EP+	FC1: 7827 FC2: 7036	20	0.6	10	31 ± 2	100	100, 30, 10, 3.3, 1.1, nM in duplicate	210	600	30	30	3M $MgCl_2$	1:1	BiaEval 2.0
T36	T100	CM5	HBS-EP+	FC1: blank FC2: 8074	20	0.6	20	234 ± 2	100	10, 3.3, 1.1, 0.37 nM in duplicate	210	10 nM curves at 1200; all others at 600	30	30	3M $MgCl_2$	1:1	BiaEval 2.0
T88	T100	CM5	HBS-EP+	FC1: 7827 FC4: 6936	20	0.6	20	82 ± 2	100	10, 3.3, 1.1, 0.37 nM in duplicate	210	600	30	30	3M $MgCl_2$	1:1	BiaEval 2.0
T93	T100	CM5	HBS-EP+	FC1: blank FC2: 8074	20	0.6	20	231 ± 3	100	10, 3.3, 1.1, 0.37, 0.12 nM in duplicate	210	10 nM curves at 1200; all others at 600	30	30	3M $MgCl_2$	1:1	BiaEval 2.0
T117	3000	CM5	HBS-EP + 0.1 mg/mL BSA	FC1: 15704 FC2: 15461	10	0.5	90	122 ± 2	50	10, 3.3, 1.1, 0.37, 0.12, 0.041 nM in duplicate Additional 10 nM set in triplicate	250	Conc series at 250; 10 nM set at 3 hours	50	30	10 mM Glycine, pH 1.5	1:1	BiaEval 4.1
T161	3000	CM5	HBS-EP + 0.1 mg/mL BSA	FC1: 21558 FC4: 21034	10	0.5	72	128 ± 2	50	10, 3.3, 1.1, 0.37, 0.12, 0.041 nM in duplicate	250	1200	50	30	10 mM Glycine, pH 1.5	1:1	BiaEval 4.1
T176	T100	CM5	HBS-EP+	FC1: 7827 FC3: 7939	20	0.6	15	74 ± 3	100	10, 3.3, 1.1, 0.123, 0.041 nM in duplicate; single cycle 0.37 nM	210	10 nM curves at 1200; all others at 600	30	30	3M $MgCl_2$	1:1	BiaEval 2.0
T278	T100	CM5	HBS-EP+	FC1: 7827 FC2: 7036	20	0.6	10	32 ± 2	100	30, 10, 3.3, 1.1 nM in duplicate	210	1 30 nM curve at 1200; all others at 600	30	30	3M $MgCl_2$	1:1	BiaEval 2.0

2.7.2 Electron density maps

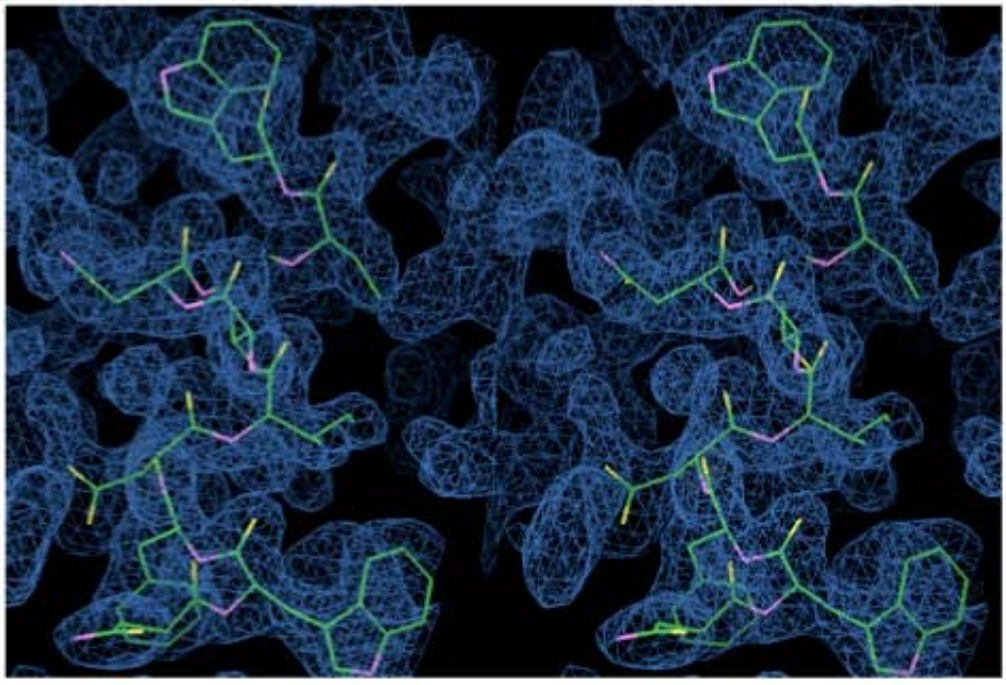


Continue I

C



D



Continue II

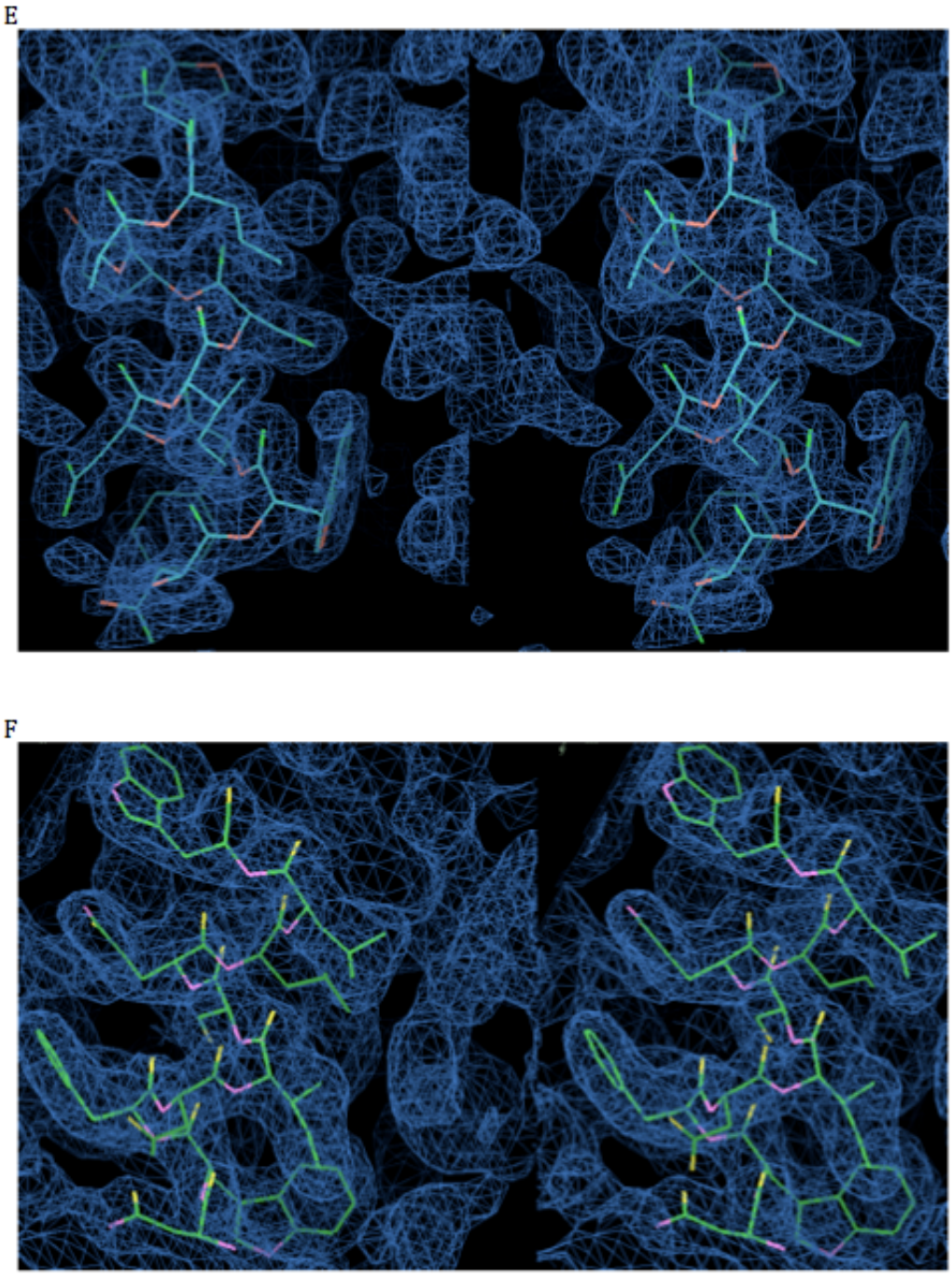


Figure 2.11: Stereo images of weighted $2F_{obs}$-F_{calc} electron density maps. Electron density maps are shown as implemented in Refmac [84], contoured at 1.0σ, for the section corresponding to the epitope, superimposed on the final structure (colored by atom type; chain A shown in all cases). (A) residues 138-147 (GWFDITAYLW) in T18; (B) residues 15-24 (NWFDITGILW) in T88 (complexed with Fv); (C) residues 87-89 (NWFDISQSLW) in T93 (complexed with Fv); (D) residues 119-128 (NWFDITNVLW) in T117; (E) residues 77-86 (NWFDISQALW) in T161; (F) residues 43-52 (NWFDISQLLW) for T246. In all frames, the N-terminus is at the bottom of the figure; in the some cases, residues could not be confidently modeled because of the quality of the density and have been omitted from the final structures. Figures were generated with Coot [83].

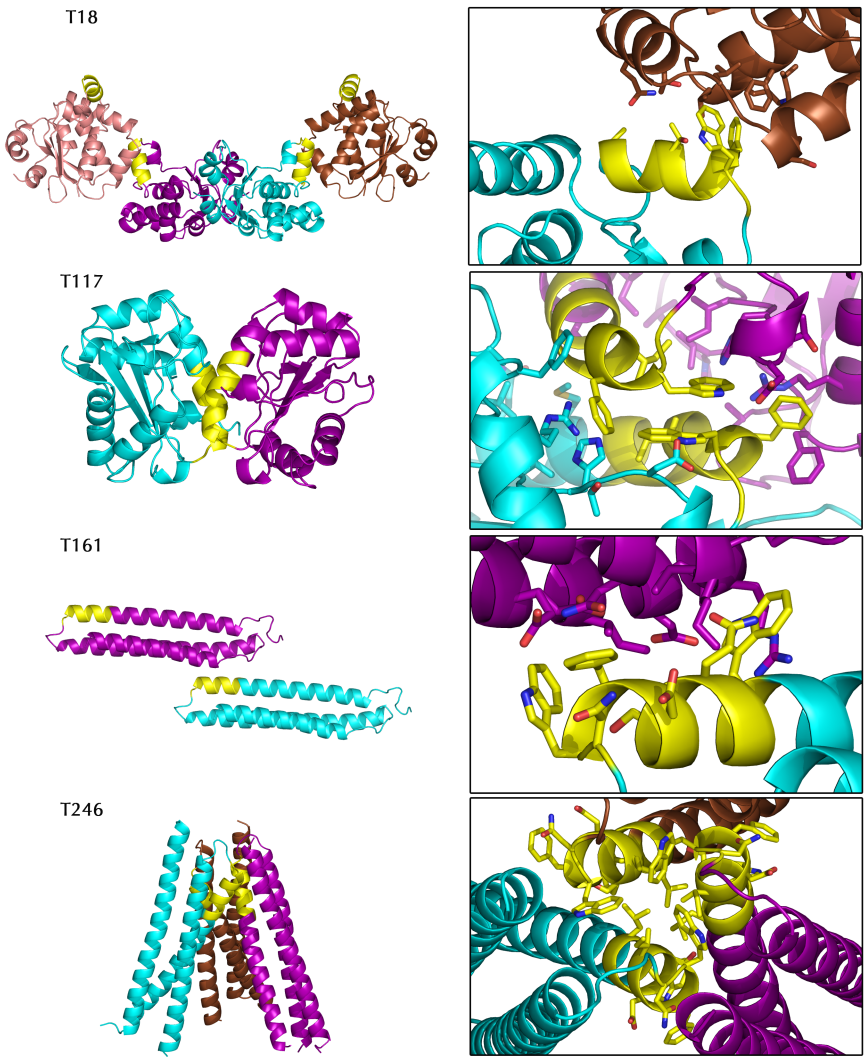


Figure 2.12: Epitope side chain conformations are influenced by protein-protein interactions within crystals of unliganded epitope-scaffolds. Shown on the left are epitope-scaffold multimers in the crystal asymmetric unit (T117 dimer, T246 trimer) or epitope-scaffolds with selected nearest neighbors generated using crystallographic symmetry (T18 dimer, T161 monomer). Shown on the right are zoomed-in and rotated views of the protein-protein interactions involving the epitope, with side-chains shown in stick if they are part of the epitope or are within 4.5 Å of the epitope. In all panels the epitope is shown in yellow. T18 is dimeric in solution; the putative structure of the T18 dimer (cyan and purple) buries 706 $Å^2$ per monomer and exposes two epitopes, consistent with measured 2:2 stoichimetry. Symmetry neighbors (brown, salmon) pack against the epitope on each monomer in the dimer (burying 337 $Å^2$ on the cyan subunit and 275 $Å^2$ on the brown subunit). T117 and T246 are dimeric and trimeric in solution, respectively; both asymmetric units contain the appropriate number of chains and bury sufficient area (1231 $Å^2$ per monomer by T117 and 1049 $Å^2$ per monomer by T246) to explain the solution multimer state. In both T117 and T246 multimers, the epitope is buried at the multimer interface, influencing epitope sidechain conformations (different subunits colored in cyan, purple and brown). T161 is a monomer in solution; the epitope packs against a neighbor (purple) in the crystal.

Table 2.8: Hydrogen bonds and salt bridge interactions. The hydrogen bonds between the complexes were assigned using the hbond application included in the CCP4 software package (v.6.1)[82].

2FX7			T88-Fv			T93-Fv		
Peptide Atom	Fab Atom	Distance (Å)	Epitope-Scaffold Atom	Fv Atom	Distance (Å)	Epitope-Scaffold Atom	Fv Atom	Distance (Å)
						Ile S86–O	Ser L95–OG	2.68
Asn P671–ND2	Tyr L91–O	2.90				Asn S87–ND2	Tyr L92-O	3.24
Trp P672–N	Ser L94–OG	3.10	Trp S16–N	Ser L95–OG	3.22	Trp S88–NE1	Ile H57–O	3.08
Trp P672–NE1	Ile H56-O	3.44	Trp S16–NE1	Ile H57– 0	3.34	Trp S88–N	Ser L95–OG	2.88
						Asp S90–OD1	Lys L33-NZ	2.75
Thr P676–OG1	Glu H95–OE1	2.69	Thr S20–OG1	Glu H99-OE1	2.75	Ser S92–OG	Glu H99-OE1	2.98
Thr P676–OG1	Glu H95–OE2	3.3[a]						
Trp P680–NE1	Leu H100–O	3.6[a]	Trp S24–NE1	Leu H107–O	3.12			
			Asp V37–OD1	Glu L99–OE1	2.70			
			Gln V39–NE2	Gln L9 –OE1	2.77			
			Ala V40–O	Ser L95–OG	2.75			

[a] Interactions listed in Cardoso et al. [61] not found in our analysis.

Table 2.9: Van der Waals contacts in 2FX7 and epitope-scaffolds in complex with the 4E10 Fv. The interactions were assessed using the contact application included in the CCP4 software package(v. 6.1)[82].

2FX7		T88-Fv		T93-Fv	
Peptide	4E10 Fab	Epitope-Scaffold	4E10 Fv	Epitope-Scaffold	4E10 Fv
				Asn S74	Leu(H54), Leu(H55)
				Asp S75	Leu(H55)
				Ala S78	Ile(H57)
				Lys S84	Ser(L95)
				Ile S86	Ser(L95)
Asn P671	Tyr(L91), Gly(L92), Gln(L93), Ser(L94)	Asn S15	Gly(L93), Ser(L95), Gln(L94)	Asn S87	Ser(L95)
Trp P672	Ser(L94), Ala(H33), Gly(H50), Val(H51), Ile(H52), Ile(H56), Asn(H58)	Trp S16	Ser(L95), Ile(H57), Asn(H59), Ala(H33), Gly(H50), Val(H51), Ile(H52)	Trp S88	Ser(L95), Ile(H57), Thr(H58), Asn(H59), Ala(H33), Gly(H50), Val(H51), Ile(H52)
Phe P673	Tyr(L91), Gln(L93), Ser(L94), Ser(L96), Trp(H47)	Phe S17	Tyr(L92), Gln(L94), Ser(L95), Ser(L97), Trp(H47)	Phe S89	Tyr(L92), Gln(L94), Ser(L95), Ser(L97), Trp(H47), Asn(H59)
Asp P674	Lys(L32)	Asp S18	-	Asp S90	Lys(L33)
Trp P675	Ile(H52),Leu(H54), Ile(H56)	Ile S19	Ile(H52)	Ile S91	Ile(H52)
Thr P676	Thr(H31),Glu(H95), Pro(H100)	Thr S20	Thr(H31), Ala(H33), Glu(H99), Pro(H110)	Ser S92	Ile(H52), Glu(H99), Pro(H110)
Asn P677	Lys(H100), Pro(H100)	Gly S21	Pro(H110)	Gln S93	Lys(H109), Pro(H110)
Leu P679	a	Leu S23	Leu(H54)	Leu S95	Ile(H52), Leu(H54)
Trp P680	Leu(H100), Gly(H100), Lys(H100)	Trp S24	Leu(H107), Lys(H109), Gly(H108), Pro(H110)	Trp S96	Tyr(H32), Leu(H107), Lys(H109), Gly(H108), Pro(H110)
Tyr P681	Lys(H100)	-	-		
		Ser S36	Ile(H57)		
		Asp V37	Ala(L0)		
		Val V38	Ala(L0)[a]		
		Val V39	Gln(L94), Ala(L0)[a]		
		Ala V40	Ser(L95)		
		Val V42	Ile(H57)		
		Asn V63	Asn(L31)a		
		Ser V64	Asn(L31)		
		Gly V65	Asn(L31)		

[a] Interactions listed in Cardoso et al. [61] not found in our analysis.

Chapter 3

COMPUTATIONAL PROTEIN DESIGN USING FLEXIBLE BACKBONE REMODELING AND RESURFACING: CASE STUDIES IN STRUCTURE-BASED ANTIGEN DESIGN

3.1 Abstract

Computational protein design has promise for vaccine design and other applications. We previously transplanted the HIV 4E10 epitope onto non-HIV protein scaffolds for structural stabilization and immune presentation. Here we developed two methods to optimize the structure of an antigen flexible backbone remodeling and resurfacing and we applied those methods to a 4E10 scaffold. In flexible backbone remodeling, an existing backbone segment is replaced by a *de novo* designed segment of pre-specified length and secondary structure. With remodeling we replaced a potentially immunodominant domain on the scaffold with a helix-loop segment that made intimate contact to the protein core. All three domain trim designs tested experimentally had improved thermal stability and similar binding affinity for the 4E10 antibody compared to the parent scaffold. A crystal structure of one design had 0.8 Å backbone RMSD to the computational model in the rebuilt region. Comparison of parent and trimmed scaffold reactivity to anti-parent sera confirmed the deletion of an immunodominant domain. In resurfacing, the surface of an antigen outside a target epitope is redesigned to obtain variants that maintain only the target epitope. Resurfaced variants of two scaffolds were designed in which 50 positions amounting to 40% of the protein sequences were mutated. Surface patch analyses indicated that most potential antibody footprints outside the 4E10 epitope were altered. The resurfaced variants maintained thermal stability and binding affinity. These results indicate that flexible backbone remodeling and resurfacing are useful tools for antigen optimization and protein engineering generally.

3.2 Introduction

Improvements in our ability to manipulate protein structure by design may facilitate development of therapeutic molecules such as vaccines or minimized binding proteins, of industrially useful molecules such as enzymes, and of structurally-tailored reagents for research. Computational protein design using "fixed backbone" methods, in which amino acid sequences are designed to stabilize pre-existing backbone conformations, has had many successes including: thermo-stabilization[96, 97, 98, 99] and hyper-solubilization[100] of proteins, design of metal[6] and small molecule binding proteins[101], novel enzymes [31, 30], enhancing or switching specificities at protein-protein[26, 24, 102, 103, 104], protein-peptide[27] and protein-DNA[28] interfaces, enhancing antibody[105] and T cell receptor[106] affinities for their targets, and stabilizing structural epitopes for immune presentation (Correia et al., in press; Ofek et al., in press).

More advanced "flexible backbone" design methods that manipulate protein backbone structure have even greater potential, as in principle they will allow protein designers maximum freedom to tailor protein structure and function to the needs of their applications. "Flexible backbone" methods (reviewed in[107]) search both backbone conformation and sequence space for low energy designs and have been employed less commonly than "fixed backbone" methods, owing to the increased computational cost and difficulties in developing appropriate sampling algorithms and accurate scoring functions. Nevertheless, several design achievements have been made. Su et al. and Harbury et al. treated secondary structural elements as rigid objects and varied their relative distances and orientations through algebraic supersecondary structure parametrization[108, 19, 19]. This approach, though limited to folds that can be parametrized, led to the design of novel right-handed coiled-coils, and the crystal structure of a tetrameric right-handed coiled-coil closely matched the design model[19]. Kuhlman et al. pioneered a more general method for flexible backbone design that involved iterating between fragment-based structure prediction and sequence design[21]. The authors used this method to design a 93 residue α/β protein with a novel fold; the protein was extremely thermostable and the model agreed with the crystal structure with a C-alpha RMSD of 1.2 Å. A similar iterative strategy was employed to build and

design three 10-residue loops onto the surface of tenascin, using loop modeling instead of full protein structure prediction[48]. All three proteins were stable, and the crystal structure of one design showed close (sub-Å C-alpha RMSD) agreement between the designed and crystallized loop conformations. Finally, the specificity of an enzyme was altered by introducing a new active-site side chain in an appropriate position and then remodeling the loop connecting that side chain to the rest of the protein using a combination of constrained loop conformational sampling and sequence design[49]. Here too, crystallographic analysis revealed excellent agreement (∼1 Å C-alpha RMSD) between designed and crystallized loop conformations.

In the area of vaccine design, we recently presented a side chain grafting method for transplanting contiguous conformational epitopes onto protein scaffolds for structural stabilization and immune presentation (Figure 3.1). We demonstrated with two different epitopes that computationally designed epitope-scaffolds can (i) stabilize epitope backbone conformations with sub-Å backbone RMSD accuracy (ii) bind target antibodies with considerably higher affinity than cognate peptide and (iii) elicit antibodies specific for the structure of the transplanted epitope. Aside from their use as vaccine candidates, such epitope-scaffolds may be useful tools to test various hypotheses concerning the relationship between structure and immunogenicity.

Here, we sought to modify the structure of one such epitope-scaffold in two different ways domain trimming by flexible backbone remodeling (Figure 3.1) and resurfacing with the aim of significantly altering the structure and antigenic surface outside our epitope of interest without harming the thermal stability, solubility, or high affinity for cognate antibody. The long-term aim is to test whether molecules with such modifications can be employed to heighten the immune reponse to the transplanted epitope.

Domain trimming was employed to eliminate potentially immunodominant, off-target epitopes and required flexible backbone design. Here we present an efficient and generalizable flexible backbone design procedure that allows one to trim a domain or segment and in its place build and design a new segment of pre-specified length and secondary structure(s). To evaluate the performance of the protocol, three different trimmed constructs were designed and experimentally characterized, and one crystal structure was determined.

A detailed structural comparison of the design model versus crystal structure demonstrated excellent agreement overall but also revealed subtle structural inaccuracies in the design model; additional computational analysis was undertaken to understand these inaccuracies and suggest improvements to the computational protocol. Finally, we tested the hypothesis that domain trimming would eliminate distracting epitopes, by comparing the reactivity of untrimmed and trimmed scaffolds to guinea pig sera elicited by the untrimmed scaffold.

Resurfacing was employed to redesign the surface of the epitope-scaffold outside the epitope of interest without degrading stability, solubility, and high-affinity for the cognate antibody. The purpose of resurfacing was to develop pairs of molecules with minimal antigenic cross-reactivity except for the epitope of interest (here the 4E10 epitope); if successful this approach could be expanded to develop larger sets (n-mers rather than pairs) of resurfaced variants; development of such sets of minimally but specifically cross-reactive antigens would allow future cocktail and prime-boost immunization experiments to test whether resurfacing can be employed to focus immune responses to a particular structural epitope. We have recently reported the design of resurfaced HIV gp120 molecules for use as reagents to isolate novel HIV-neutralizing antibodies with unusual potency and breadth of cross-reactivity[109]. In that work, the resurfacing design protocol took advantage of evolutionary information that may not be available for most molecules. Here we present a simpler and more general resurfacing protocol that favors mutations for amino acids with different physical and chemical properties, and test it on an epitope-scaffold protein. Two resurfaced variants were designed that modify approximately 40% of the overall protein sequence; a computational surface patch analysis was developed to quantify the degree of antigenic modification, and the stability, multimeric state, and binding affinity of the molecules was assessed.

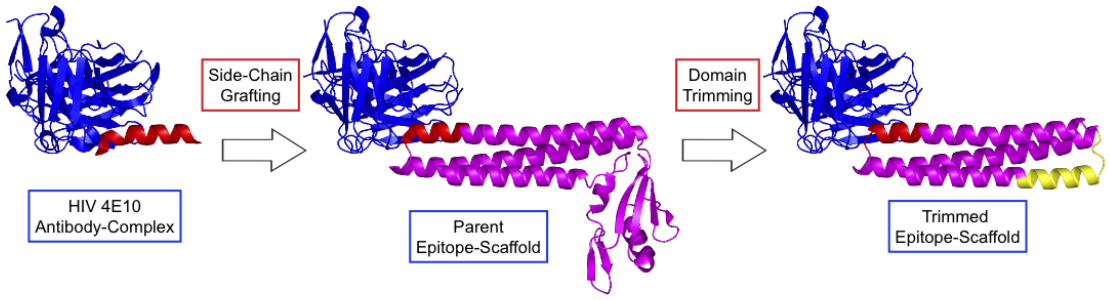


Figure 3.1: Design stages of a trimmed epitope-scaffold.The crystal structure of the 4E10 epitope in complex with the antibody (PDB ID: 2FX7) was the starting point for the design of epitope-scaffolds using the side chain grafting procedure. The initial epitope-scaffold was further optimized by trimming a domain unnecessary for epitope recognition.

3.3 Materials and Methods

3.3.1 Flexible backbone remodeling

The computational protocol to rebuild and design *de novo* backbone segments were comprised of five stages. In stage (i) length and secondary structure were defined for the *de novo* segment. In this particular problem, 10 different combinations of helix/loop lengths were attempted. In stage (ii) 500 new backbone segments were generated using an *ab initio* fragment insertion protocol as described by Simon et al. [22]. The fragments used for the loop building stage were selected wholly based upon secondary structure assignment; no sequence information was considered. To maintain proper polypeptide chain connectivity, cyclic-coordinate descent (CCD) chain closure[110] was used within the rebuilt region[38, 39]. Residue identities at the positions within helix and loop secondary structure were pre-defined as alanine and glycine, respectively. Each new configuration of the rebuilt loop was evaluated according to a Metropolis criterion based on the total energy of the protein computed with a low resolution energy function[111]. The final combinations of secondary structure linker lengths were selected in order to exclude linkers presenting kinked helices or extra long loop segments. From the 10 initial combinations of helix/loop lengths, two different segment lengths were selected to proceed on the design stage: length 16 with secondary structure (HHHHHHHHHHHHLLLL), and length 17 with one additional loop residue yielding secondary structure (HHHHHHHHHHHHLLLLL). In stage (iii) the remodeled segment and neighboring residues were designed with standard RosettaDesign[15] in order to find sequences that maximally stabilize the conformation of the newly remodeled region. For design positions, all amino acids except cysteine were allowed during the design process. The best 3 scoring designs from each of the initial backbones were relaxed using a full-atom energy function[111]. In stage (iv), filtering based on structural features was performed. From the resulting 1500 designs output in stage (iii), all designs with buried polar atoms were dismissed. This filtering resulted in 423 and 59 designs remaining for the linker lengths 16 and 17, respectively. A second filtering step was performed using a double criteria where all the designs ranked within the best 20% by full-atom energy and packing score[112] were selected. Final designs for experimental characterization were selected from

the remaining pool, and were the two best ranking designs by full-atom energy for linker length 16 and the single best ranking design for linker length 17. In the final stage (v) to optimize solubility, a few hydrophobics in surface patches were manually redesigned to alanine, lysine, or arginine. The Rosetta design algorithm is not yet optimized for solubility so restricted design on the surfaces of designed proteins is often necessary. The manual mutations were as follows: T72 (I36A, I37A, I40A, Y41K), T81 (A44R), T93 (L33K, I36A, L37A, I40K).

3.3.2 Structure prediction

The computational model for the trimmed design T93 was the starting point for the loop relax simulations[39]. NNMAKE fragments[111] were generated with the secondary structure assignment observed in the actual crystal structure of T93. The segment for the prediction calculations was 27 residues long, comprising residue 31 to 56 as per crystal structure numbering. 20000 models were generated and compared with the crystal structure in terms of structural and energetic convergence. Two different types of refinement (CCD-refine and all-atom relax) were performed on the 100 lowest energy structures generated by the loop relax simulation, to sample low energy conformations more finely. CCD-refine[110], which is essentially the loop modeling procedure described in [39] but lacking the fragment insertion stage, sampled new conformations for a section of the protein around the rebuilt region. All-atom relax[111] allowed small moves in backbone torsion angles and cycles of repacking followed by minimization of side chain rotamers over the entire protein. The relaxed crystal structure ensemble was generated with the all-atom relax procedure.

3.3.3 Resurfacing

Surface residues were selected based on visual inspection excluding the epitope-antibody interface. The resurfacing design calculations were carried with standard RosettaDesign[15]. The amino acids allowed in each of the selected surface positions were ADEGKMNQRST. 1000 models were generated for each starting structure and the final designs to be experimentally characterized were selected according to a sequence difference score (SEQ_DIFF),

in which preference was given to designs with least degree of sequence identity to the starting sequence. The sequence difference score was composed of two terms and it was computed as follows: SEQ_DIFF = FLIP_SCORE + (0.5 x NEG_POL_SCORE). The amino acids were grouped according to their chemical properties: negatives (D and E), positives (R and K), polars (N and Q) and other (A, F, G, H, I, L, M, P, S, T, V, W, Y). The FLIP_SCORE was 1 in case there was one of two scenarios: a) sequence change from negative or polar to positive or vice-versa; b) sequence change from group 'other' to any of the remaining groups or vice-versa. The NEG_POL_SCORE was 1 if there was a sequence change from negative to polar or vice-versa.

3.3.4 Surface patch analysis

The surface patch analysis was performed by placing spheres centered on each of the exposed C-alpha atoms of the crystal structure or computational models under analysis. A patch was composed of all solvent exposed residues with C-alpha atoms contained within the sphere. By definition, any given residue was allowed to belong to multiple patches. Residues were defined as solvent exposed if more than 30% of the side chain surface area was exposed to solvent according to NACCESS[92]. The sphere radius was 11 Å, yielding an area of 380 $Å^2$ on the largest cross section of the sphere. This cross section aimed to be a low-resolution approximation of a minimal antibody footprint. Once patches were defined in an initial molecule, each patch was compared to its corresponding patch in the resurfaced variant and the percentage of mutations within each given patch was computed.

3.3.5 Experimental methods

Protein expression, purification, and biophysical characterization were carried out as described in Correia et al. Experimental conditions for the SPR experiments are presented in Table 3.3. ELISA experiments were performed as described in Correia et al.

3.4 Results

3.4.1 Computational flexible backbone remodeling

To heighten the immune response to a particular epitope on an antigen, one rational strategy is to minimize the size of the antigen and eliminate other potentially immunogenic domains. Previously, we designed a panel of epitope-scaffolds for the conserved HIV 4E10 epitope by the side chain grafting method (Correia et al., in press), with the goal of eliciting antibodies specific for the structurally stabilized 4E10 epitope. One particular epitope-scaffold was based on a mutant of the E. coli Ribosome Recycling Factor (RRF) (PDB code: 1ISE, chain A) and was a two-domain protein with the conserved RRF fold[113] that includes a three-helix bundle and a globular beta-alpha-beta sandwich, with an overall "L" shape. This epitope-scaffold (T36) had an extremely high affinity for the monoclonal antibody 4E10 (Table 3.1), so we decided to minimize its size in an effort to optimize the immunogenicity of the transplanted epitope. We had transplanted the 4E10 epitope to one end of the three-helix bundle, and we sought to remove the globular domain without compromising the thermal stability or 4E10 binding affinity of the molecule. In particular, we aimed to replace the alpha-beta globular domain with a single helix followed by a short loop, in which the helix-loop segment would make contacts with the protein core promoting structural integrity and stability.

To meet this goal, we devised an efficient and general computational protocol involving flexible backbone design to remove the extraneous domain and remodel the region with segments of pre-specified length and secondary structure. Briefly, the protocol (Figure 3.2) was composed of six main stages: (i) selection of one or more lengths for the new segment and specification of the secondary structure at each position in the segment, (ii) *de novo* building of the segment, where backbone conformations were sampled by insertion of fragments that satisfy the desired secondary structure(s), (iii) sequence design in which low energy amino acid sequences were selected for each backbone conformation, (iv) all-atom relaxation in which each structure was optimized with small backbone perturbations and side chain rotamer packing and minimization, (v) filtering in which designs were selected based on buried unsatisfied polar atoms, packing and Rosetta full-atom energy, and (vi)

solubility optimization in which hydrophobic residues in surface patches were mutated to polars or alanine. Protocol details are given in Materials and Methods. Three designs (T72, T81, and T93) were selected for experimental characterization (Figure 3.3); two had a *de novo* segment with length 16 (T72 and T93) and the third had length 17 with one more residue in the loop connection (T81). The selected computational designs had significantly different sequences in the remodeled segment and in neighboring positions (Table 3.2 and Figure 3.3).

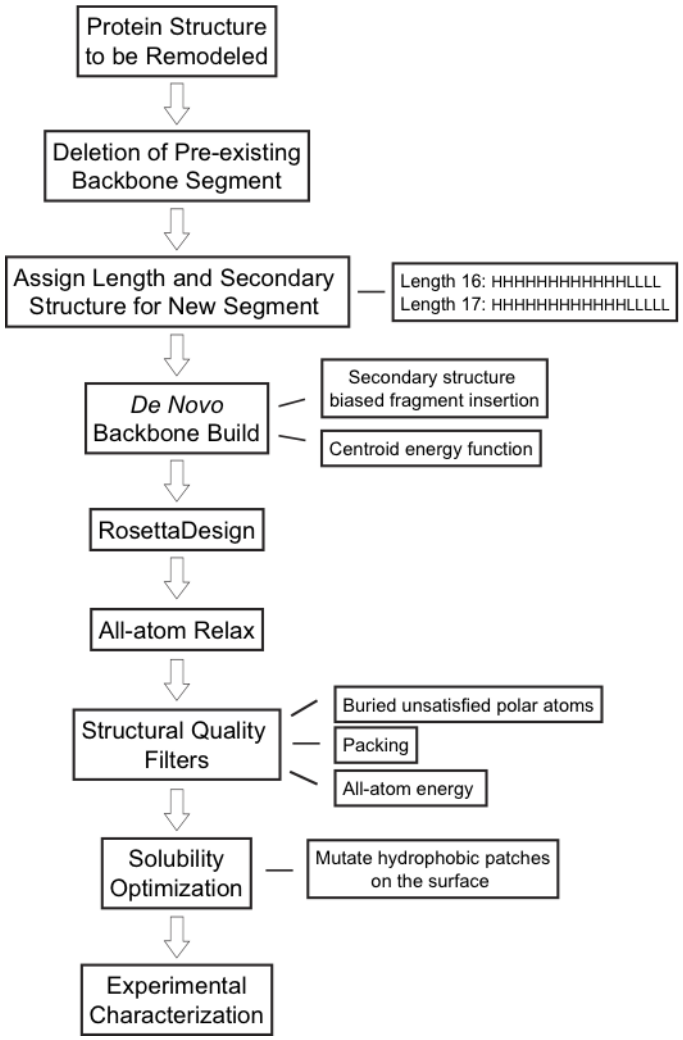


Figure 3.2: Overview of the computational protocol to build and design new segments of backbone.

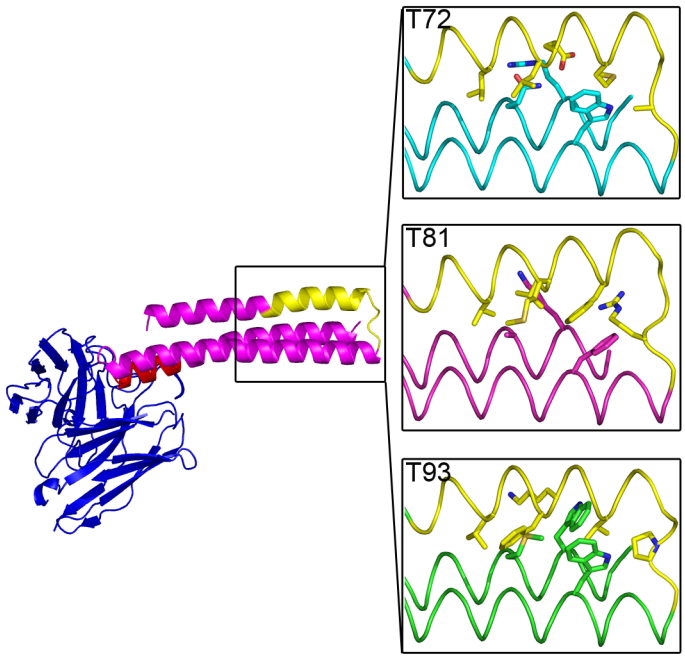


Figure 3.3: Computational models of trimmed scaffolds selected for experimental characterization. Backbone is shown in tube representation and core side chains of the redesigned region are shown in sticks representation.

Table 3.1: Characterization of the solution properties of the trimmed and resurfaced epitope-scaffolds.

Epitope-scaffold	Solution Multimer State	Tm (°C)	Fv:epitope-scaffold Binding Stoichiometry	Interaction Parameters (SPR)					
				Analyte: 4E10 Fv Ligand: epitope-scaffold[1]			Analyte: epitope-scaffold Ligand: 4E10 IgG[2]		
				k_{OFF}/k_{ON} (pM)	k_{ON} ($M^{-1}s^{-1}$)	k_{OFF} (s^{-1})	k_{OFF}/k_{ON} (pM)	k_{ON} ($M^{-1}s^{-1}$)	k_{OFF} (s^{-1})
T36 (parent)	monomer	44	1:1	14.2	5.89×10^{6}	8.34×10^{-5}	44	1.95×10^{6}	8.59×10^{-5}
T72	monomer	70	1:1	228	1.86×10^{6}	4.24×10^{-4}	17.7	1.15×10^{7}	2.03×10^{-4}
T81	monomer	79	1:1	16.3	4.07×10^{6}	6.61×10^{-5}	33	2.82×10^{6}	9.30×10^{-5}
T93	monomer	79	1:1	7.49	6.98×10^{6}	5.23×10^{-5}	27.5	3.17×10^{6}	8.73×10^{-5}
T345	monomer/ dimer	84	1:1	50.6	1.57×10^{6}	7.95×10^{-5}	16.2	2.39×10^{6}	3.88×10^{-5}
T353	monomer/ dimer	>95	1:1	522	1.52×10^{6}	7.96×10^{-4}	85.4	3.48×10^{6}	2.97×10^{-4}
T355	monomer	90	1:1	165	1.36×10^{6}	2.54×10^{-4}	22.4	3.70×10^{6}	8.28×10^{-5}

[1] Epitope-scaffold was immobilized by biotinylation and capture on SA chip.

[2] 4E10 IgG was captured by amine-coupled anti-human IgG.

For all reported values, standard error is $\leq \pm 3$ of the last significant figure.

Table 3.2: Structural and sequence features of selected computational designs selected for experimental characterization. Residues in the core are colored in red.

Design	Helix-loop linker[a]	Sequence
T72	12-4	[1-20]QK**L**KQ**LE**AA**M**RAKG**A**-TEER...W..A[47-108]Q..R...G
T81	12-5	[1-20]EQ**V**RR**ML**ER**F**ESLQT**R**SWFD...F..I[48-109]A..K...-
T93	12-4	[1-20]NK**L**DK**FK**AA**V**RKVF**P**-TEER...W..I[47-108]M..W...G

[a] Lengths off the secondary structure elements used in the computational procedure

3.4.2 Biophysical characterization of the computational designs

Design of well-folded, soluble, stable and functional proteins based on flexible backbone design remains a major challenge due to inaccuracies in energy functions and difficulties in sampling protein conformational spaces. To evaluate the performance of the flexible backbone protocol presented here, we subjected several of our computationally designed proteins to biophysical characterization. Solution multimeric states were determined using size exclusion chromatography (SEC) and static light scattering (SLS), thermal melting points were assessed by circular dichroism (CD) melting analysis, and dissociation constants (K_D) and kinetic constants (k_{on}, k_{off}) for antibody binding were determined using SPR (see Materials and Methods). This biophysical data is collected in Table 3.1 for the parent scaffold (T36) and the three trimmed scaffolds illustrated in Figure 3.3 (T72, T81, and T93).

All three domain trimmed variants (T72, T81, T93) were monodisperse and monomeric according to SEC-SLS analysis, as was the parent (T36). The thermal stabilities of the domain trimmed scaffolds (T72, T81 and T93) were 70, 79, and 79 ^oC, respectively, considerably higher than the parent with $T_m = 44\ ^oC$ (Figure 3.4).

Dissociation and kinetic constants, were evaluated by SPR in two formats: (I) epitope-scaffold amine-coupled to the SPR chip and 4E10 scFv flowing as analyte; (II) epitope-scaffold as analyte and 4E10 IgG captured by amine-coupled anti-human IgG. SPR sensorgrams are shown in Figure 3.5. The two formats produced similar results in that nearly all the individual K_D, k_{on}, and k_{off} fit values agreed within a factor of 3 measured in one format or the other. The parent scaffold (T36) bound 4E10 very tightly, with $K_D = 14 \pm 3$ and 44 ± 3 pM as measured in formats I and II, respectively. Two of the three trimmed scaffolds (T81 and T93) bound 4E10 with similar or marginally better dissociation constants than the parent, with a range of $K_D = 27.5$ to 33 pM in both formats for T81 and T93. The ranges of on-rates (k_{on} = (2.82 - 6.98) x 10^6 M^{-1} s^{-1}) and off-rates (k_{off} = (5.23 - 9.30) x 10^{-5} s^{-1}) for these two trimmed scaffolds were similar to the parent ranges (k_{on} = (1.95 - 5.89) x 10^6 M^{-1} s^{-1} and k_{of}f = (5.23 - 9.30) x 10_{-5} s^{-1}). The other trimmed scaffold (T72) had similar K_D as the parent scaffold when assessed in format II (K_D = 17.7 pM),

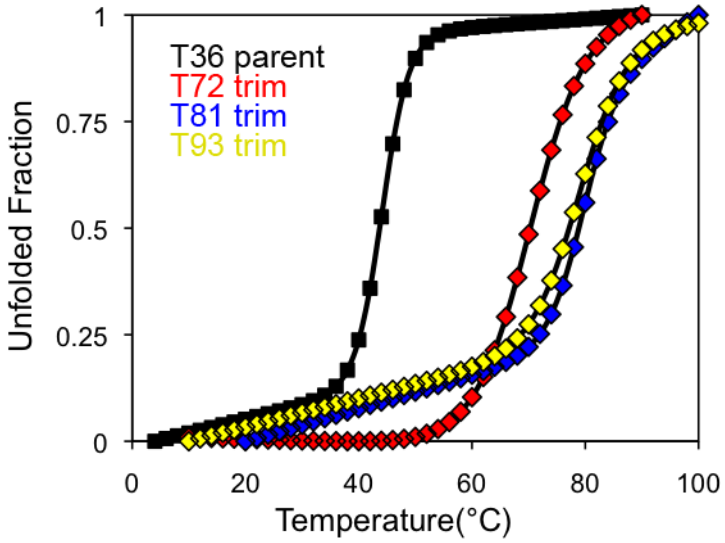


Figure 3.4: Computational designs showed improved thermal stabilities compared to the parent epitope-scaffold. Thermal stabilities were assessed by circular dichroism.

but had a slower on-rate and faster off-rate in format I compared to format II leading to a higher K_D value in format I (228 pM). The reason for this discrepancy is unclear, however the essential conclusion remains that the trimmed scaffolds generally maintained the high affinity of the parent.

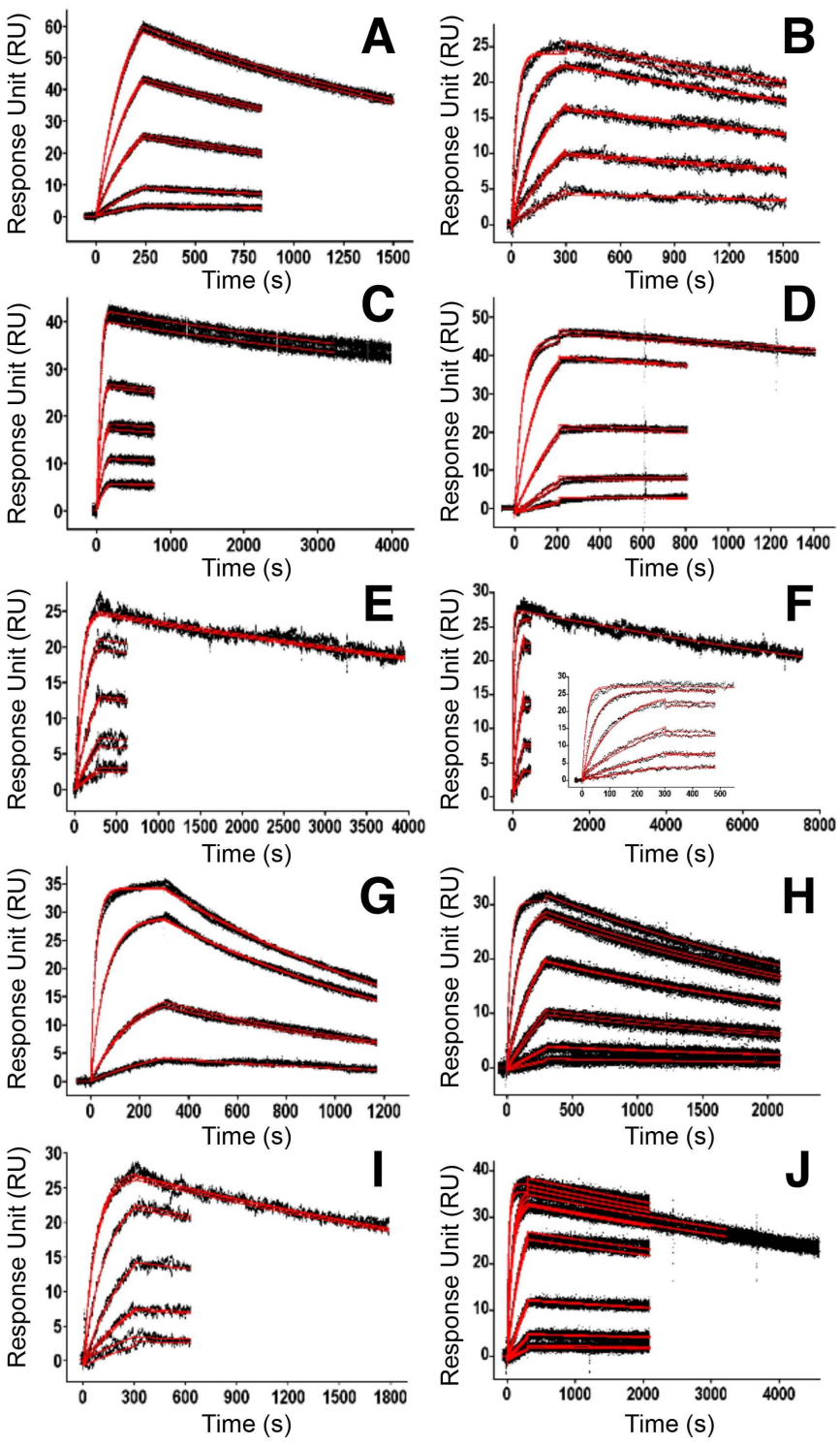


Figure 3.5: SPR binding kinetics of 4E10 Fv/IgG with epitope-scaffolds. Double reference-subtracted SPR responses are in black with the corresponding kinetic fits in red. (A) Biotin-T72 coupled to chip. (B) T72, 4E10 IgG captured. (C) Biotin-T81 coupled to the chip. (D) T81, 4E10 IgG captured. (E) Biotin-T345 coupled to the chip. (F) T345, 4E10 IgG captured. (G) Biotin-T353 coupled to the chip. (H) T353, 4E10 IgG captured. (I) Biotin-T355 coupled to the chip. (J) T355, 4E10 IgG captured.

3.4.3 Structural analysis of a computational design

To assess the atomic accuracy of the computational designs, we sought to compare our design models to crystal structures of the designs. Crystallization trials were performed on the three trimmed variants (T72, T81, and T93), both unliganded and in complex with 4E10 scFv. A crystal structure of T93 in complex with scFv 4E10 (see Figure 3.6A) was solved at 2.8 Å resolution, and a comparison of the structure and the design model in the region of the transplanted 4E10 epitope has been reported (Correia et al., in press). Here, we compare the design model and crystal structure at the other end of the three-helix bundle (Figure 3.3), in the region where the globular domain was replaced with a helix-loop segment by flexible backbone design as described above.

Overall, the T93 computational model is in close agreement with the crystal structure, with a RMSD of 1.1 Å over the backbone atoms (N, CA, C, and O) of the epitope-scaffold (Figure 3.6B). Within the remodeled segment alone (residues 27 to 45 in 3LHP numbering), the computational model and the crystal structure are also in very close agreement, with a backbone RMSD of 0.8 Å.

Side chain rotamers (chi1, chi2) were correctly predicted for 7 out of 11 non-alanine side chains that constitute the core (side chain surface area solvent accessible $< 30\%$) of the remodeled region (see examples in Figure 3.6C), but incorrectly predicted for four such side chains. Two of the incorrectly predicted side chains were errors in Leu chi2 angles that caused relatively minor alterations in packing interactions with neighboring side chains, but the other two mistakes were errors in the chi1 angle of a Phe (F34) and the chi2 angle of a Trp (W51) that caused significant differences in packing between the design and the crystal structure (Figure 3.6D).

Though the design was quite successful overall, we sought to understand the origin of these rotamer errors in order to gain insight on how to further optimize our computational protocol. Structural comparison of the design and crystal structure identified two potential aspects of the design model backbone that could have led to the incorrect rotamers for F34 and W51. First, the design model has a different backbone conformation in the loop of the remodeled region that prevents W51 from adopting the rotamer in the crystal structure a

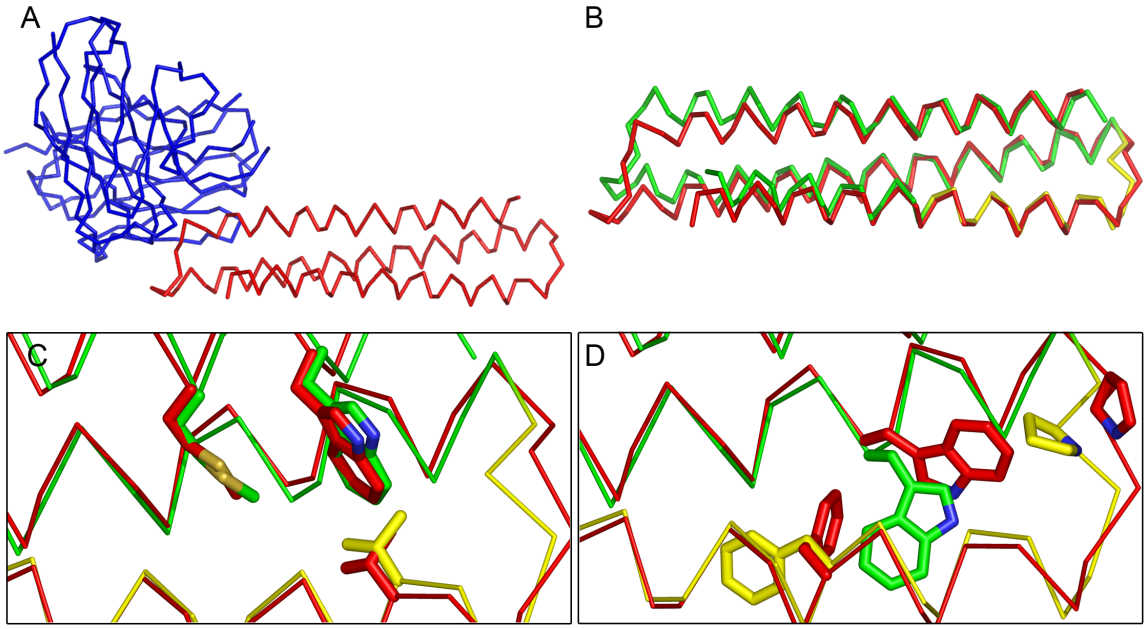


Figure 3.6: Structural comparison of the crystal structure with the computational model for the T93 trimmed epitope-scaffold. (A) Crystal structure of a trimmed epitope-scaffold in complex with the 4E10 scFv. The backbones are shown in ribbons, T93 is colored in red and the 4E10 scFv is colored in blue. (B) Backbone superposition of T93 crystal structure and computational model. T93 crystal structure is colored in red and the computational model is colored in green and yellow in the remodeled region. (C) Three examples of core side chains accurately predicted in the computational model (M116, W119, and V38). (D) The three notable core side chains inaccurately predicted in the computational model (F34, P43, and W51).

proline residue in the loop (P43) points outward in the crystal structure and allows W51 to pack into the core adjacent to P43 (Figure 3.6D), while in the design model P35 points inward and sterically prevents W51 from packing into the core (Figures 3.6D and 3.7). Second, the design model has a smaller spacing between helices one and two in the region of F34 and W51 that prevents F34 from adopting the rotamer in the crystal structure the 8.04 Å distance between the C-beta of F34 and the C-alpha of W51 in the crystal structure allows F34 to pack adjacent to W51 (Figure 3.6D) but the shorter distance (7.12 Å) in the design model sterically prevents F34 from assuming that rotamer (Figures 3.6D and 3.7). Thus, small inaccuracies in the backbone conformation (within the 0.7 Å backbone RMSD

from the crystal structure) were likely sufficient to enforce two packing errors.

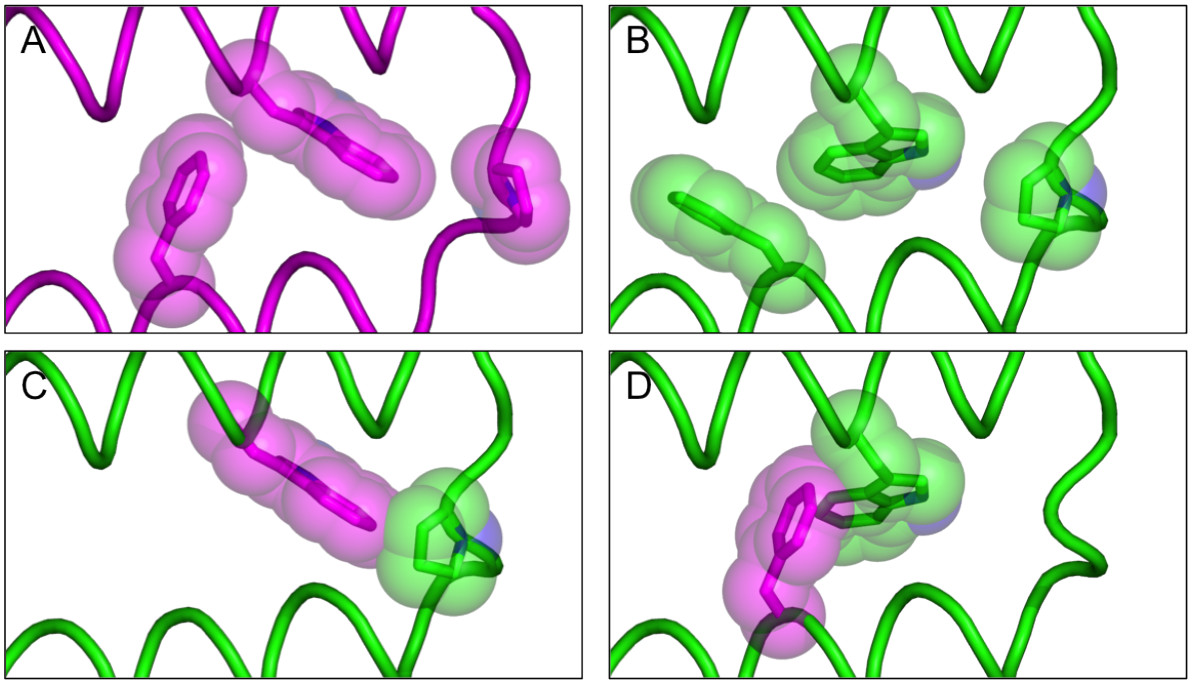


Figure 3.7: Some side chain rotamers in the T93 crystal structure are incompatible with the backbone conformation of the initial computational design model. Side chains in magenta were observed in the crystal structure and side chains in yellow and green were predicted in the computational design. (A) Crystal structure of T93 showing the F34, P43, W51 in sphere representation. (B) Computational model with side chains inaccurately predicted (F34, P43, W51) shown in sticks and spheres. (C) Native rotamer of tryptophan (W51) is incompatible with the proline (P43) conformation observed in the computational model. (D) Steric clash between the native phenylalanine (F34) and tryptophan (W51) from the design. Additionally, steric clashes occur between native F54 and both the C-alpha and C-beta atoms of the neighboring helix (helix from W51), showing that the distance or relative orientation or both are not exactly the same in the model and the crystal structure.

3.4.4 Analysis of the computational protocol

The above observations indicate that in order to improve flexible backbone computational design methods, it is important to understand the origin of even small errors such as incorrect helix-helix spacing, conformation of the proline, and the two incorrect side chain predictions discussed above. The goal of such analysis is to understand how to modify the

design procedure to avoid such inaccuracies in the future. Such errors can generally arise from imperfect energy functions, incomplete conformational sampling, or both. One hint that the errors noted above were due to incomplete sampling came from our observation that the Rosetta all-atom energy was considerably lower for the relaxed crystal structure than for the design model. This implied that the Rosetta energy function was capable of discriminating between similar backbones (0.7 Å RMSD) with only a few side chain packing differences.

To further investigate the hypothesis that the above inaccuracies were due to incomplete sampling, we used Rosetta to generate ensembles of low energy conformations for the designed scaffold and assessed whether finer sampling could recapitulate the conformation and low energy of the crystal structure. We generated three ensembles with different combinations of methods that differed in both the scope (local vs global) and amplitude of their sampling across the protein. The first ensemble was created by a Rosetta loop relax protocol [39] that aggressively sampled conformations for the remodeled region only (Figure 3.9). The second ensemble was generated by a loop refine protocol (Cyclic Coordinate Descent (CCD)[110, 39] refine) starting from the 100 lowest energy models in first ensemble, but the scope of sampling was expanded to include residues neighboring the remodeled region (Figure 3.9B). The third ensemble, with the greatest scope of sampling, also derived from the 100 lowest energy models in the first ensemble but was generated by an all-atom relaxation of side chains and backbone torsions, allowing small movements over the entire model[111, 77] (Figure 3.9C).

Remarkably, the third ensemble did produce models with low energies and conformational features (helix-helix distance, F34 and W51 rotamers, P43 conformation) closely matching the crystal structure (Figures 3.9C, 3.10 and S4), while the other ensembles never produced all of these features in a single model. This set of simulations confirms that the small errors in the original design model were indeed due to incomplete sampling, and demonstrates that the Rosetta energy function is capable of discriminating subtle structural features.

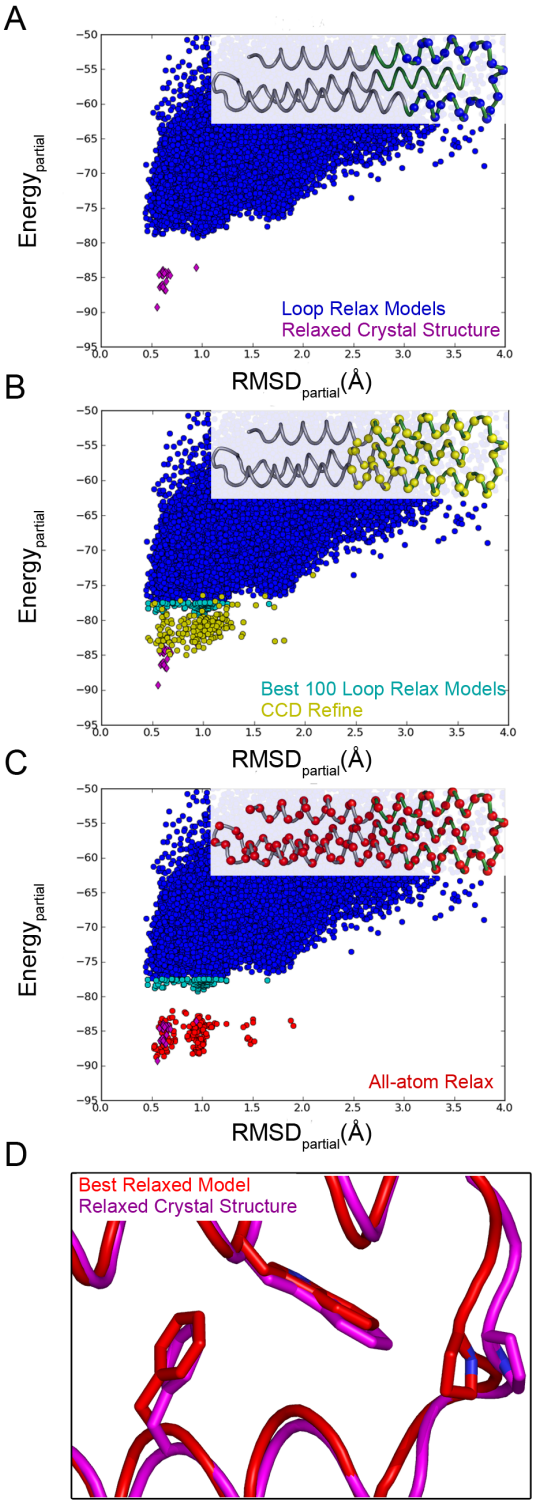


Figure 3.9: Structural prediction of the *de novo* remodeled segment in T93. $\text{RMSD}_{partial}$ is the backbone RMSD to the T93 crystal structure, computed over the remodeled region (residues 25-55 in 3LHP numbering) and at positions on the third helix in close contact to the remodeled region (residues 109-123). $\text{Energy}_{partial}$ is the full-atom energy computed over the same region where the RMSD was measured. The partial energy is computed to dismiss any structural differences caused by the co-crystallization of T93 with the 4E10 scFv. (A) Energy versus RMSD for loop relax models of the T93 design (blue) and for the all-atom relaxed T93 crystal structure (magenta). Both simulations achieve low RMSD but the relaxed crystal structure is considerably lower in energy. Inset: The moveable region for loop relax is indicated by blue spheres and the region over which $\text{RMSD}_{partial}$ was measured is indicated by green tube. The entire crystal structure was all-atom relaxed. (B) Energy versus RMSD as in (A) but also including results for CCD-refine (yellow) of the lowest energy models from loop-relax (cyan). The lowest energy CCD-refine models are similar in energy to the all-atom relaxed crystal structure. Inset: The moveable region for CCD-refine (residues 23-65 and 104-122 using 3LHP numbering, larger than the moveable region for loop relax) is indicated by yellow spheres. (C) Energy versus RMSD as in (A) but also including results for all-atom relax (red) of the lowest energy models from loop-relax (cyan). The energy distribution of these all-atom relaxed models is similar to the distribution for the all-atom relaxed crystal structure. (D) Structural superposition of a low energy and low RMSD all-atom relaxed design (red) with the lowest energy relaxed crystal structure of T93 (magenta). Side chain rotamers inaccurately predicted in the initial computational design but successfully recovered after additional conformational sampling (F34, P43, W51) are shown in sticks.

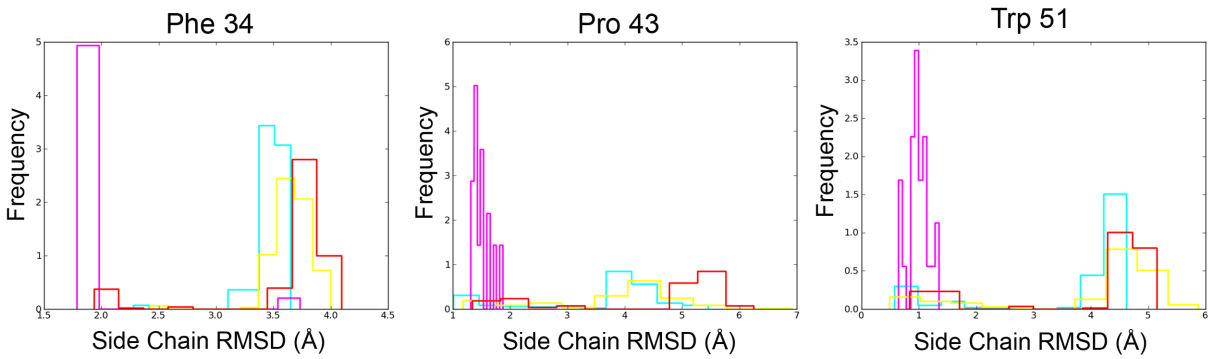


Figure 3.10: Distributions of side chain RMSDs between native rotamers in T93 crystal structure and models generated by several prediction protocols. The residues shown were inaccurately predicted in the initial computational design: F34, P43 and W51 (3LHP numbering). Methods used to generate models are described in the Material and Methods section.

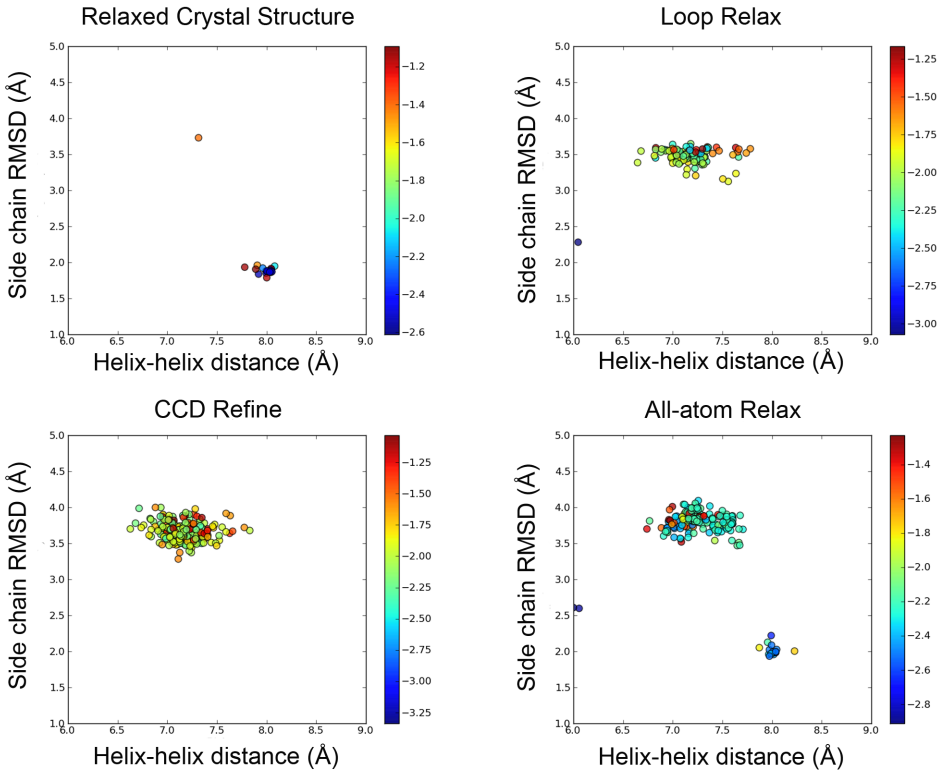


Figure 3.11: Dependency between helix distance and side chain conformation. The x-axis shows the distance between helix 1 (C-beta, F34) and helix 2 (C-alpha, W51) on the bundle; the y-axis shows the side chain RMSD of F34 between the native rotamer and the rotamers observed in models generated by different prediction-refinement protocols. The color scale applied to the dots represents the energy of the residue F34 according to Rosettas full atom scoring function. Overall, the different simulations show that the correct rotamer (low RMSD on the y axis) is only sampled if the helix-helix distance is very close to 8 Å. When the correct conformation is sampled it is recognized as an energy minimum.

3.4.5 Domain trimming eliminates "distracting" epitopes

Having confirmed the biophysical properties of three trimmed scaffolds (T72 ,T81, and T93) and confirmed the atomic structure of one (T93), we sought to test our original hypothesis that the globular domain of the parent scaffold T36 was immunodominant. For this test, we compared the binding of the guinea pig T36 antisera to T36 and to one of the trimmed scaffolds. As shown in Figure 3.12, the T36 antisera from all three of the T36-immunized animals reacted strongly with T36 itself, but the reactivity to the trimmed scaffold T72 was significantly (4.5-fold) decreased.

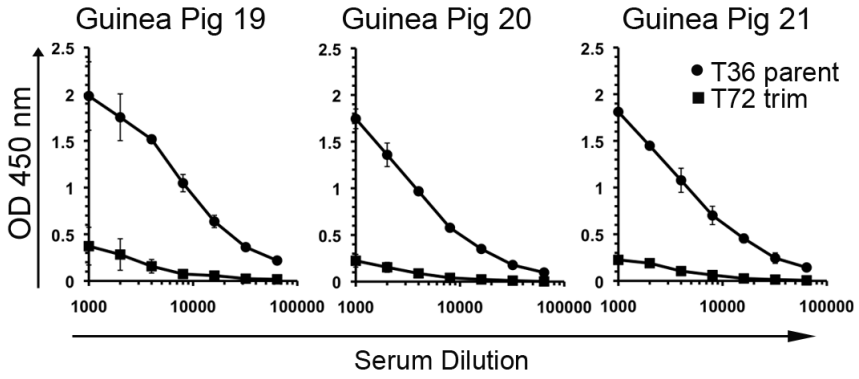


Figure 3.12: ELISA reactivity of the trimmed T93 scaffold to guinea pig sera elicited by the parent epitope-scaffold T36. The pre-bleed measurements were subtracted to account for the endogenous response of the animals.

3.4.6 Resurfacing

Above we have addressed flexible backbone remodeling of an antigen - a 4E10 epitope scaffold - in order to remove a potentially immunodominant domain. Here, we address a different method in an attempt to alter the antigenic properties of a given molecule, namely resurfacing. Resurfacing aims to change most or all of the exposed residues on the surface of a protein, keeping only a functional site or epitope of interest unchanged. Potential applications of resurfacing might include (A) improving solubility or stability, or altering the multimerization state of a protein, or (B) developing variants of a given

antigen that contain no cross-reactive epitopes except for a particular epitope of interest, or both. The achievement of (B) in particular could provide molecular reagents useful for probing antibody specificities in sera or isolating monoclonal antibodies with particular specificities, and could also provide opportunities for novel immunization schemes involving cocktails or heterologous prime-boost of resurfaced antigens. Epitope-scaffolds are useful model systems to investigate this strategy since they are small, stable and bind with high affinities to the target antibodies. In terms of design difficulty, resurfacing is less aggressive than remodeling/domain trimming, because resurfacing requires only fixed backbone design of surface-exposed side chains; however, altering a large fraction of a protein surface implies mutating a significant fraction of the protein sequence and can potentially impact expression, stability, solubility and binding affinity.

Lawrence et al. previously reported a method to redesign the surface of proteins to increase their solubility and stability by greatly increasing the net charge ("super charging")[100]. T72, T81, and T93 were neutral or slightly positively charged (net charges +2, 0, and +2, respectively). As a test of super-charging, we redesigned the surface of T72 to reach large positive (+26) and negative (-22) net charges, but neither of those variants could be purified. Based on those results we decided to limit our resurfacing designs to maintain similar net charge.

As a first step to evaluate the viability of the resurfacing strategy, T93 and a further remodeled variant of T93 (T345) were resurfaced while retaining the 4E10 epitope, and experimentally characterized. T345 was designed using flexible backbone remodeling to both shorten a loop leading into the 4E10 epitope and to lengthen a helix underneath the epitope, both aspects of the design attempting to further stabilize the antibody-bound conformation of the epitope. T345 had marginally improved stability over T93, and similar binding affinity and kinetics to T93, but was a mixed monomer/dimer in solution whereas T93 was a pure monomer (see Table 3.1).

T93 and T345 surfaces outside the 4E10 epitope were mutated using RosettaDesign in conjunction with a customized score to favor mutations to amino acids with distinct chemical properties and maximize the difference in terms of the number of side chain atoms (see methods). The resulting "resurfaced" epitope-scaffolds were T353 (derived from T93),

and T355 (derived from T345). T353 and T355 had 49 and 48 surface residue mutations, respectively, meaning that 42% and 40% of each protein was mutated (Figure 3.13). In both the designs approximately 45% of the mutations were from charged to charged amino-acids, 25% were apolar to charged, and roughly 15% were charged to polar (Figure S5). The theoretical net charge of the resurfaced designs was not allowed to deviate far from T93; the theoretical net charges of T353 and T355 were +2 and +3, respectively.

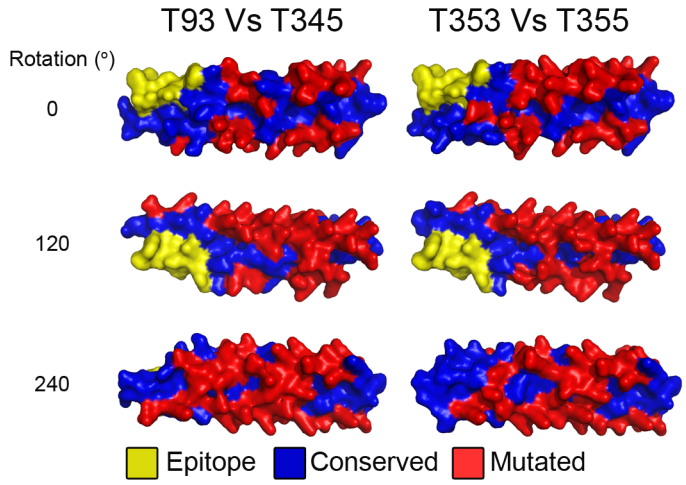


Figure 3.13: Mutations on the resurfaced variants of the epitope-scaffolds. Epitope is colored in yellow, the unchanged residues are colored in blue and the surface mutated residues are shown in red.

A surface patch analysis was performed to quantify the degree of antigenic change that was imposed on the molecules by resurfacing. The surface of each parent molecule was divided into patches and then compared to the corresponding patches on resurfaced variants. Each patch was defined as the solvent-exposed amino acids enclosed within a sphere of radius 11 Å centered on the C-alpha atom of a surface-exposed residue; patches were defined for spheres centered on every surface-exposed position (Figure 3.15A). The sphere radius was chosen so that the enclosed surface area (∼400 Å^2) would approximate a minimal antibody binding footprint [114]. The aim of the patch analysis was to assess the number of mutations in each of the patches and identify regions of the surface that might require additional mutations. Ideally only patches centered on epitope residues would have few or

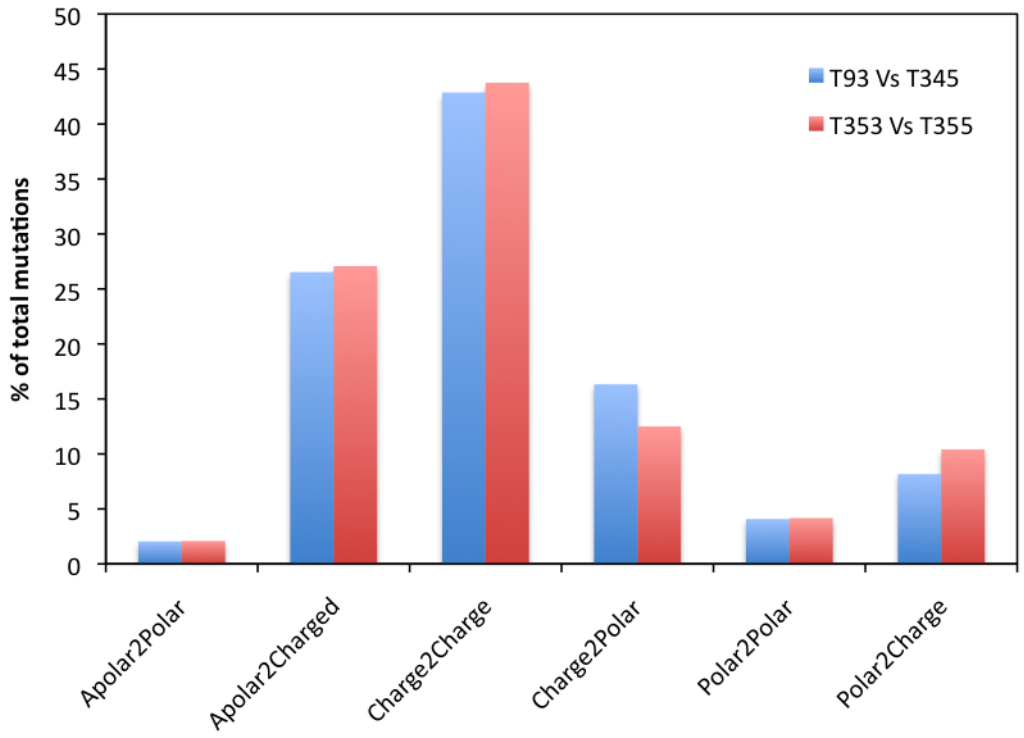


Figure 3.14: Categorization of the types of mutations performed in the Resurfacing procedure. The two most prevalent types of mutations employed were mutations from apolar to charged amino acids and from charged to another charged amino acid.

no mutations in the resurfaced variants compared to the parent molecules (Figure 3.15 - see sphere enclosing yellow epitope). For T353, the resurfaced variant of T93, 48 out of 78 patches presented 50% or less of sequence identity, and the statistics were approximately the same for T355 and its parent T345 (Figure 3.15B,C).

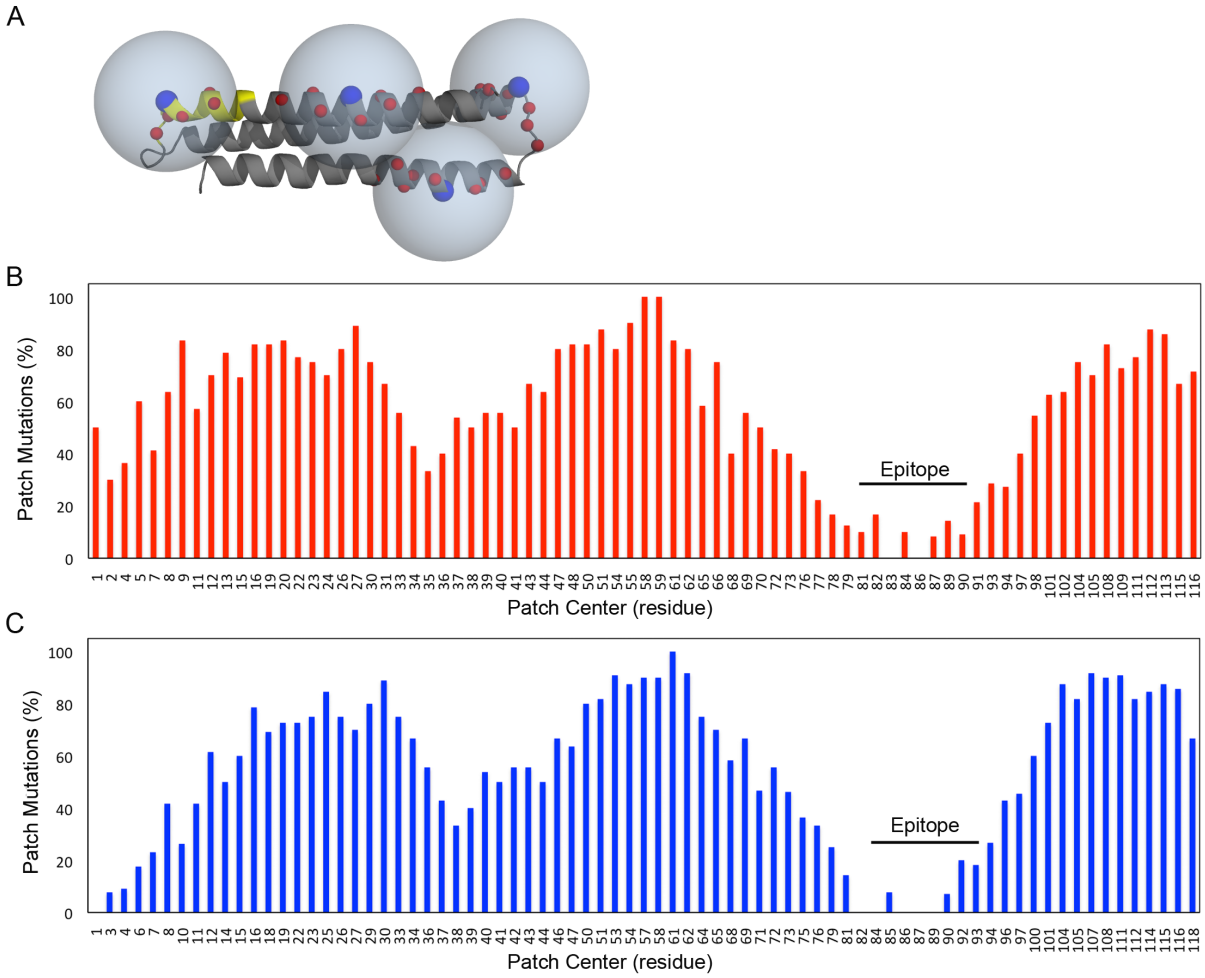


Figure 3.15: Surface patch analysis. (A) Schematic of the surface patch analysis procedure. Spheres in light blue define patches of surface residues that are then compared with the corresponding patch on the resurfaced variant; the number of mutations performed at each patch is computed. Dark blue spheres represent examples of exposed residues that were centers of patches and red spheres represent residues that belong to a given patch. Only residues with side chain solvent exposure greater then 30% (according to NACCESS) were considered as part of a patch. (B) Percent of residues mutated within each patch of the resurfaced variant T345 relative to its initial design T93. (C) Percent of residues mutated within each patch of the resurfaced variant T355 relative to its initial design T353.

The biophysical properties of the resurfaced designs were generally consistent with the parent molecules. The resurfaced designs had improved thermal stabilities, with T353 having a $T_m > 95$ ^{o}C compared to $T_m = 79$ ^{o}C for the parent T93, and T355 having a $T_m = 90$ ^{o} compared to 84 ^{o} for the parent T345. The effect of resurfacing on multimeric state was mixed; T353 was a mixed monomer/dimer in solution even though the parent T93 was a pure monomer; on the other hand T355 was a pure monomer in solution while the parent T345 was a mixed monomer/dimer. We do know that the 4E10 epitope itself can mediate dimerization but that other scaffold-scaffold interactions can modulate this tendency (Correia et al., in press, Xu et al. in preparation), and so it is possible that these changes in multimer state are due to such modulations in the propensity for epitope-mediated dimerization. The resurfaced designs maintained high affinity for 4E10, though the dissociation constant for T353 was increased 70-fold ($K_D = 522$ pM compared to 7.49 pM for the parent T93) and the dissociation constant for T355 was increased 3-fold ($K_D =$ 165 pM compared to 50.6 pM for the parent T345).

3.5 Discussion

Here we presented an efficient and general method for remodeling of segments of a protein, in which an existing segment is replaced by a *de novo* designed segment of pre-specified secondary structure. Given a high-resolution protein structure and a segment that is to be replaced, the method entails pre-specifying the length and secondary structure of the new segment, *de novo* building an ensemble of backbone conformations for the segment in the new protein, fixed backbone design of the segment and neighboring positions in each member of the ensemble, and filtering the resulting models by various energetic and structural criteria. We also presented a simple method for resurfacing a protein, in which the entire exposed surface of a protein outside a functional site is subjected to redesign. We tested these methods on an epitope-scaffold, a scaffold protein onto which we had previously transplanted the HIV 4E10 conserved neutralization epitope, and we presented biophysical and structural analysis of remodeled and resurfaced variants of the initial epitope-scaffold.

In our test of the remodeling method, the initial scaffold had the "L"-shaped RRF fold with a three-helix bundle and an internal 77 residue globular domain connected to the

bundle by two flexible linkers; by *de novo* design, we replaced the globular domain with a 16 or 17 residue helix-loop linker that made intimate contact to the core of the remaining three-helix bundle. Three designs with significantly different sequences were purified from *E. coli*; all had markedly improved thermal stability over the parent scaffold (Tm improved from 44 ^{o}C to 70-79 ^{o}C), all were monomeric in solution like the parent, and all had similar or marginally improved binding affinity for the 4E10 antibody. Crystallographic analysis of one of the trimmed scaffolds in complex with 4E10 demonstrated close agreement between the design model and the crystal structure in the region that was remodeled (backbone RMSD = 0.8 Å); the remodeling procedure is therefore capable of atomic-level accuracy in design of novel protein segments. Finally, we demonstrated that the globular domain deletion from the scaffold succeeded to eliminate an immunodominant domain, because polyclonal sera elicited by the parent scaffold cross-reacted only weakly to the trimmed scaffold.

Despite the overall success of this design, our comparison of the crystal structure to its corresponding design model revealed several subtle differences and our follow-up analysis (a) showed that the Rosetta energy function recognized the crystal structure as lower in energy than the design model and (b) implied that increased conformational sampling in our design protocol would have avoided the observed inaccuracies. There are two aspects of the protocol that we will improve. First, the initial generation of backbone conformations using the Rosetta centroid energy function is susceptible to building overly compact protein structures[115, 111]; here we may have accentuated that problem by using alanine centroids during *de novo* linker building. Our observation of overly tight helix-helix spacing (and the effect of that tight spacing on side chain packing) within the remodeled region highlights that this aspect of the protocol should be improved. Second, the success of the full-atom optimization protocol in recapitulating key structural features of the crystal structure absent in the design model indicates that addition of a full-atom optimization stage in between the current design and filtering stages would be beneficial.

Our tests of the resurfacing method started with two of the trimmed scaffolds that we had designed by remodeling. Resurfacing was achieved by fixed-backbone protein design of surface exposed residues using RosettaDesign and selecting design sequences with the least

similarity to the parent as judged by a "sequence difference" score. Two resurfaced proteins were designed, one for each of two parent trimmed scaffolds, in which 50 residues (40% of the protein sequences) were mutated. A patch analysis was developed to assess whether all potential antibody epitopes had been altered; patch analysis of the two resurfaced variants indicated that most but not all epitopes had been heavily modified. The resurfaced proteins had moderately improved thermal stabilities but also larger dissociation constants for the target antibody 4E10.

These results indicate that *de novo* remodeling, in particular trimming, can be employed to minimize therapeutic proteins or other proteins of interest without sacrificing stability, solubility, or functional activity (e.g. binding). While trimming of loops and domains has long been achieved by addition of short flexible linkers (for example the use of "GAG" in place of the variable loops on HIV gp120[116], or the use of "GGGG" in place of the RRF globular domain[117, 118]) those efforts typically require a certain amount of empirical testing and are not suited to design segments that interact with and stabilize the core of the protein as we have done here. Our results on resurfacing indicate that it may be possible to modify all potential antibody epitopes (or other types of binding sites) on the surface of a protein outside a particular function or binding site, while maintaining the stability, solubility, and functional activity (binding affinity) of the protein of interest. This resurfacing technology may have applications to improve solubility and/or alter multimerization propensity; to simplify the binding profiles of proteins with multiple binding sites/partners facilitating the study of individual interactions, and to design novel antigens and serological reagents[109].

3.6 Acknowledgements

We thank Paul Murphy and Possu Huang for advice on secondary structure specific fragment insertion and David Baker for guidance and support. This work was funded by the Bill and Melinda Gates Foundation Collaboration for AIDS Vaccine Discovery (CAVD) and Bruno Emanuel Correia was supported by a fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/32958/2006).

William Ray Schief (W.R.S.) conceived and supervised the study. Bruno Emanuel Cor-

reia (B.E.C.) and W.R.S. devised the trimming protocol; B.E.C. designed trimmed scaffolds and Yih-En Ban (Y.-E.B.) designed resurfaced scaffolds. Della Friend(D.F.), Hengyu Xu, Tyler Bradley-Hewitt(T.B.-H.), Jessica F. Bruhn-Johannsen produced proteins and performed SLS and SEC analysis; Erica Boni(E.B.) and D.F. performed SPR analysis; E.B. and T.B.-H. performed CD melt analysis; B.E.C. and W.R.S. analyzed crystal structures; Katharine Ellingson performed ELISA; B.E.C. and W.R.S. wrote the paper; all authors commented on the manuscript.

3.7 Supplemental Material

3.7.1 Surface plasmon resonance experimental protocol

Table 3.3: SPR methods for 4E10 Fv analyte binding to captured biotinylated epitope-scaffolds and epitope-scaffold analyte binding to captured 4E10 IgG (shaded section).

	Biacore Instrument	Chip Type	Running Buffer	Immobilized Anti-IgG Level (RU)	Captured Biotinylated Epitope Scaffold or 4E10 IgG Ligand				4E10 Fv or Epitope Scaffold Analyte				Regeneration			Analysis	
					Flow Rate (µL/min)	Concentration (µg/mL)	Injection Time (s)	Captured Level (RU)	Flow Rate (µL/min)	Concentration	Injection Time (s)	Dissociation Time (s)	Flow Rate (µL/min)	Injection Time (s)	Regeneration Buffer	Model	Software
T72	T100	SA	HBS-EP+	NA	10	≈10	42	60	50	5, 3.75, 2.5, 1.25, 0.62 nM in duplicate	240	5 nM at 1300; all others at 600	50	5	10 mM glycine pH 1.5	1:1	BiaEval 2.0
T81	T100	SA	HBS-EP+	NA	10	≈10	12	45	50	10, 7.5, 5, 3.75, 2.5 nM in duplicate	180	10 nM at 4000; all others at 600	50	5	10 mM glycine pH 1.5	1:1	BiaEval 2.0
T345	3000	SA	HBS-EP	NA	10	≈0.3	72	25	50	5, 2.5, 1.25, 0.625 nM in duplicate 10 nM in triplicate	300	10 nM at 3660; all others at 300	50	12	10 mM glycine pH 1.5	1:1	BiaEval 4.1
T353	T100	SA	HBS-EP+	NA	10	≈1.2	36	25	50	30, 10, 3.3, 1.1 nM in duplicate	300	900	50	10	10 mM glycine pH 1.5	1:1	BiaEval 2.0
T355	3000	SA	HBS-EP	NA	10	≈0.3	102	20	50	5, 2.5, 1.25, 0.625 nM in duplicate 10 nM in triplicate	300	10 nM at 1500; all others at 300	50	12	10 mM glycine pH 1.5	1:1	BiaEval 4.1
T72	3000	CM5	HBS-EP +0.1 mg/mL BSA	FC1: 15703 FC2: 15461	10	0.5	90	124 ± 3	50	3.3, 1.1, 0.37, 0.12 in triplicate 0.04 nM in duplicate	300	1200	50	30	10 mM glycine pH 1.5	1:1	BiaEval 4.1
T81	T100	CM5	HBS-EP+	FC1: blank FC2: 8074	20	0.6	20	235 ± 5	100	10, 3.3, 1.1, 0.37, 0.12 nM in duplicate	210	10 nM at 1200; all others at 600	50	30	3 M $MgCl_2$	1:1	BiaEval 2.0
T345	3000	CM5	HBS-EP +0.1 mg/mL BSA	FC1: 21558 FC2: 21034	10	0.5	120	134 ± 4	50	30,10, 3.3, 1.1, 0.37, 0.12 nM in duplicate	300	30 nM at 7200; all others at 180	50	30	10 mM glycine pH 1.5	1:1	BiaEval 4.1
T353	T100	CM5	HBS-EP+ +0.1 mg/mL BSA	FC1: 18478 FC2: 17463	10	0.5	120	178 ± 2	50	3.3, 1.1, 0.37, 0.12. 0.04 nM in duplicate Single cycle at 10 nM	300	1800	50	30	10 mM glycine pH 1.5	1:1	BiaEval 2.0
T355	T100	CM5	HBS-EP+ +0.1 mg/mL BSA	FC1: 18478 FC2: 17463	10	0.5	120	171 ± 3	50	10, 3.3, 1.1, 0.37, 0.12. 0.04 nM in duplicate Additional duplicate 3.3 nM cycles for long diss	300	1 pair 3.3 nM at 4200; all others at 1800	50	30	10 mM glycine pH 1.5	1:1	BiaEval 2.0

Chapter 4

TRANSPLANTATION OF A COMPLEX BINDING SITE USING COMPUTATIONAL DESIGN AND *IN VITRO* EVOLUTION

4.1 Abstract

The transplantation of a complex epitope or binding site from one protein of known structure to other scaffold proteins serves as a challenging test for flexible backbone protein design and could be employed to devise novel vaccines or binding reagents. We sought to graft a complex antibody epitope involving two non-contiguous segments from the HIV gp120 envelope protein onto a heterologous scaffold protein. Within Rosetta, we developed a general computational method called Multigraft for transplantation of complex epitopes. Given an input structure of an antibody-epitope complex, Multigraft searches the PDB for suitable acceptor scaffolds using geometrical criteria, replaces the backbone of the candidate scaffolds with the epitope backbone, and employs iterative loop modeling and design to build a continuous backbone. Starting from a crystal structure of the b12 antibody in complex with gp120, a number of scaffolds were designed, produced and their affinity for b12 assayed by SPR. In that first generation, only one scaffold had a detectable and specific affinity for b12, but the K_D ($\sim$300 μM) was significantly higher than the K_D for gp120 (20 nM). To improve b12 binding, computation-guided evolution was undertaken. Cycles of conformational sampling and design were performed to diversify the sequence of the connecting regions between the epitope segments and the acceptor protein. Directed mutagenesis libraries were derived from the resulting sequences and were screened for b12 binding using yeast surface display. The best clone bound b12 with a K_D of 35 nM, comparable to the K_D for the b12-gp120 interaction. To our knowledge, this is the first successful backbone transplantation of a discontinuous binding site to a heterologous protein. Further, the approach shows how to leverage aggressive computational design with an efficient *in vitro* evolution strategy. Lastly, the optimized scaffold could be useful for HIV vaccine

research to detect or elicit b12-like antibodies.

4.2 Introduction

Several longstanding challenges have recently been met using rational protein design. Most notable achievements in protein design have relied on the atomic structure of a pre-existing backbone observed in nature. For example, the latest advances in rational enzyme design involved the selection of backbones able to support the geometry of two or three catalytic side chains [30, 31, 32]. The inclusion of backbone flexibility in design calculations presents significant challenges [119, 107], but has promise to elevate expectation of the field by allowing for more accurate protein modeling and by expanding the scope of applications of protein design.

Random mutagenesis and directed evolution approaches have become common laboratory tools to manipulate and optimize protein function. The desired improvements are typically related to binding affinities, catalytic activity and protein solution properties[120]. Some of the selection platforms have obvious advantages such as the possibility of screening large libraries without having to purify each individual sequence and allow for a reliable quantification of the property to be improved[121]. Although not required, a starting sequence with some measurable level of the desired property can contribute to the success of *in vitro* evolution experiments[30]. Frequently, multiple rounds of mutagenesis and selection are performed to generate clusters of mutations in a way that most of these mutations will synergistically improve the property under selection[122]. Computational design has been helpful in providing starting sequences[30] with large potential for improvement. Several examples have been described in the literature of computational designs successfully optimized using *in vitro* evolution and also approaches that utilize computationally inspired libraries[123, 124].

Here we describe our approach to tackle a relevant biomedical problem, the engineering of "epitope-scaffolds" onto which the complex HIV b12 broadly neutralizing epitope was transplanted. Several HIV broadly neutralizing antibodies (bNAb) have been isolated and they neutralize efficiently most strains of the virus. b12 was the first bNAb isolated from a human infected patient using phage display techniques[125]. The elicitation of bNAbs is

likely critical for the efficacy of an HIV vaccine and the retrovaccinology[50] approach has been focused in trying to elicit known bNAbs, creating the need for structure-based design approaches. In particular the epitope-scaffolds, molecules that did not evolve to escape the immune system and lack all the molecular trickery typical of the pathogen's molecules[126]. A fundamental aspect about the viability of the retrovaccinology strategy is to know if the bNAbs will efficiently neutralize the virus in humans, optimistic indications came from studies in animal models where b12 has shown sterile protection in macaques challenged with and HIV-1/SIV chimeric virus[127, 128]. Vaccines have been successful by using native molecules of pathogens or variations of live attenuated pathogens[76]. Despite all the efforts employed in the optimization of gp120 and gp41, molecules that compose the HIV viral spike, they were never able to elicit useful neutralizing responses, although several bNAbs that target both molecules are known[129]. The HIV viral spike is composed of trimers of 2 distinct subunits (gp41 and gp120); the binding surface of the broadly neutralizing antibody b12 with gp120 overlaps with the CD4 binding site on gp120, and the interaction with CD4 is conserved and required for cell-invasion[129]. Hence, the elicitation of antibodies similar to b12 would likely be relevant for vaccine development purposes.

The b12-gp120 complex has been characterized by x-ray crystallography[74], gp120 has been co-crystallized with other molecules where the binding surface also overlaps with the CD4 binding site[130, 131], showing the remarkable conformational plasticity intrinsic to this site. Additionally, the b12 binding site is a complex epitope where the interactions with gp120 are comprised within several discontinuous segments of the protein[74]. The structural knowledge accumulated over the years, enables the use of rational design strategies to transplant the key structural motif to scaffold proteins where it will be conformationally stabilized for a robust presentation to the immune system.

The transplantation of the b12 complex epitope poses a challenge to rational design strategies, serving as proxy to tackle other relevant problems using protein design. Here, we describe a general strategy to accomplish complex epitope transplantations to unrelated protein scaffolds (Figure 4.1): (i) a general matching scheme to select host scaffolds for epitope transplantation (Figure 4.2); (ii) flexible backbone design to connect the epitope to the scaffold; (iii) a structure-sequence diversification protocol (Figure 4.3) to design

computationally inspired libraries that are then screened in an *in vitro* evolution platform. To our knowledge, this is the first successful backbone transplantation of a discontinuous binding site onto an heterologous protein.

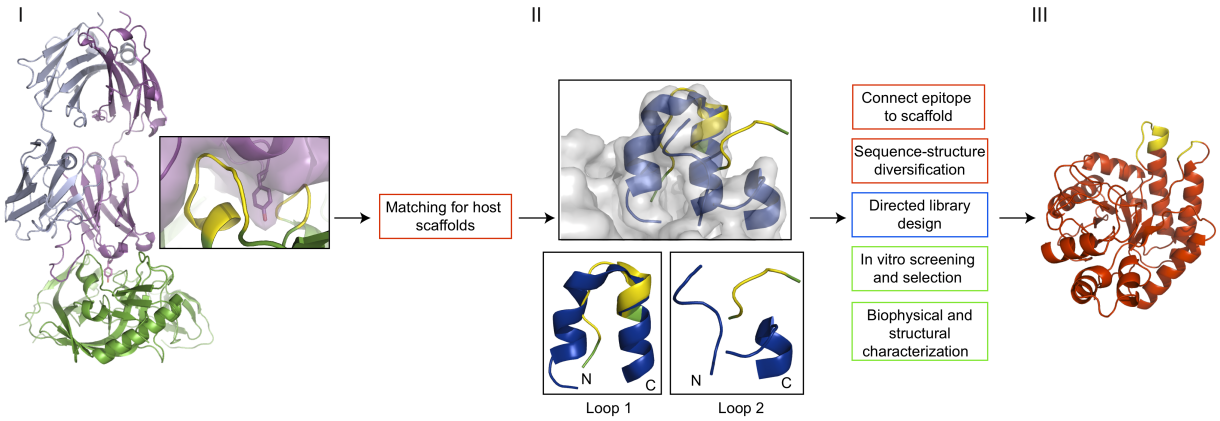


Figure 4.1: Overview of combined computational and experimental strategy. (I) b12-gp120 complex (PBDid: 2NY7) was used to identify host scaffolds. gp120 is colored in green, the b12 antibody is colored in magenta and gray, the b12 epitope is colored in yellow. (II) The matching algorithm searched for scaffolds determined the orientation of the host scaffold (blue) and the epitope loops (yellow). After a first computational design step that connects the epitope to the scaffold, computational guided libraries were designed and screened using yeast surface display. (III) Computational model of the epitope-scaffold.

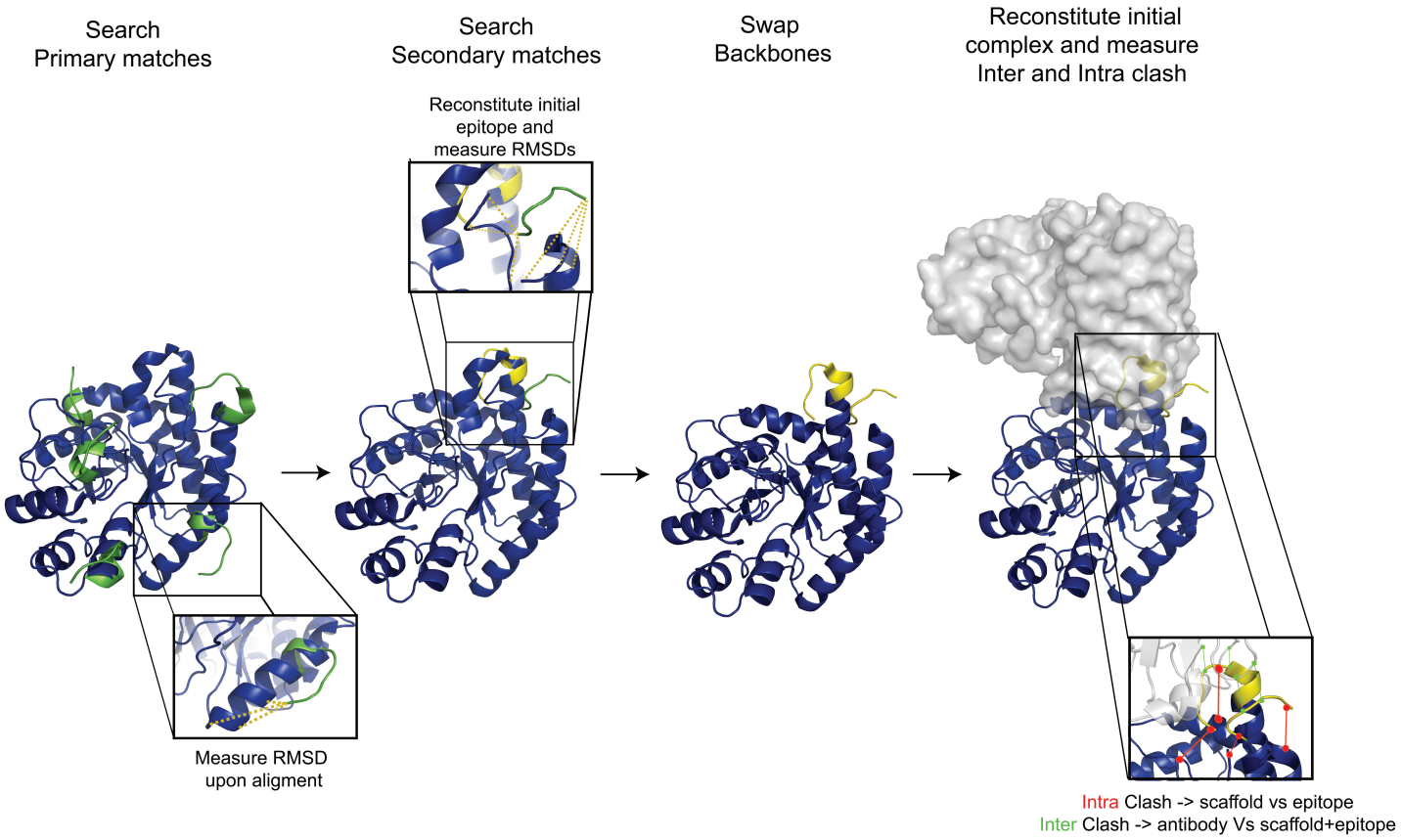


Figure 4.2: Matching algorithm to select compatible scaffolds with the epitope. The first step is to search for primary matches using one of the loops (Loop 1 colored in green and the scaffold colored in blue). The loop is aligned to all the residues in the protein and RMSDs to residues in the protein are measured if the RMSD is bellow a given threshold the match is accepted (Loop 1 colored in yellow). In the second step the original epitope is recreated by realigning Loop 2 in the correct orientation and the distance criteria is used once again to accept or reject a particular match. Once a match is found the third step is to swap the pre-existing scaffold backbone by the epitope backbone. The fourth step is to recreate the complex between epitope and antibody (showed in surface representation and colored in grey), at this stage steric clashes are measured between the epitope and the scaffold (red lines with circled edges) and also between the scaffold and the antibody (green lines with squared edges). If the scaffold is clash free, a match is accepted and carried to the design stage.

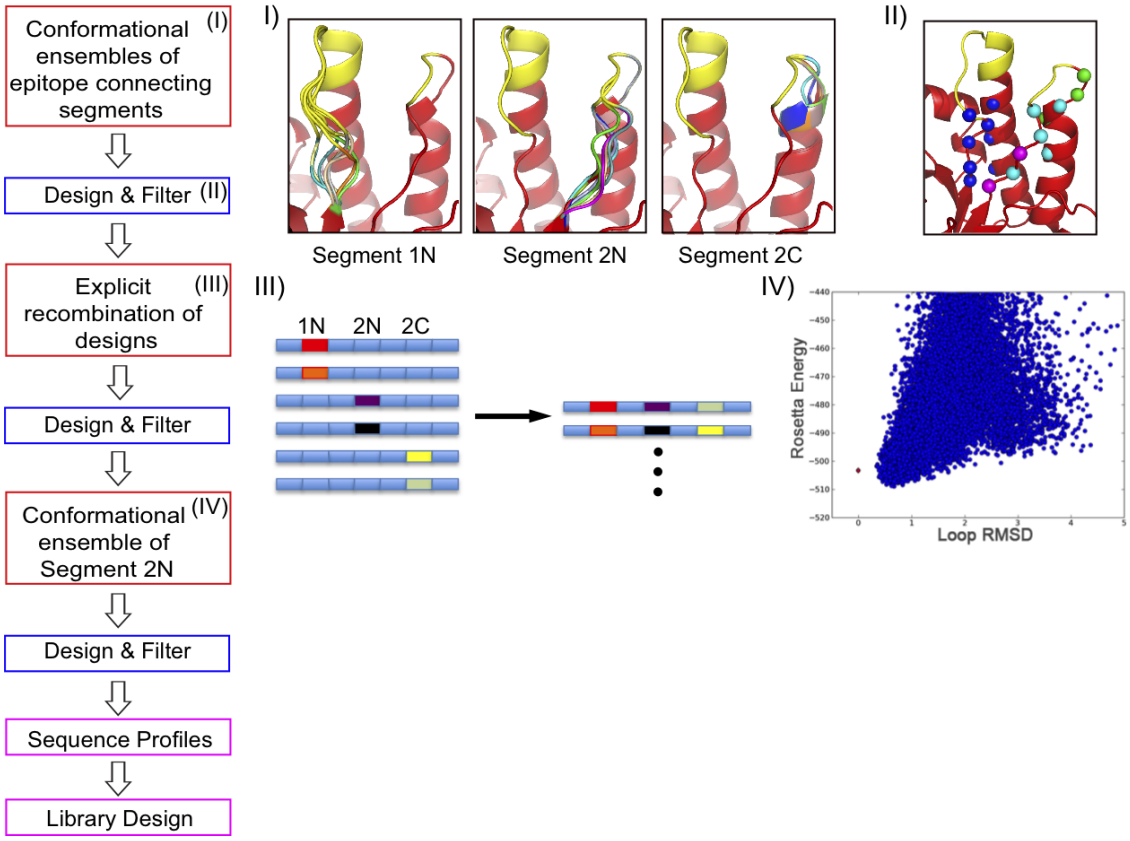


Figure 4.3: Step-by-step computational protocol for structure-sequence diversification. (I) Backbone sampling of epitope connecting segments using loop prediction type techniques[39, 110]. (II) C-alphas of design positions for each of the simulations are represented in spheres. Blue spheres were designable positions in the simulation for segment 1N, cyan and magenta positions were designed in the simulation for segment 2N and cyan and green were used on simulations of segment 2C. (III) *In silico* recombination of conformations of different segments was performed to ensure structural compatibility. (IV) After the recombination and design steps, loop prediction simulations for segment 2N were performed and the loop sequence showed energetic preference for the initial conformation of the design. Three directed libraries were designed based on the sequence profiles after filtering the designs using structural quality indicators like full-atom energy, packing score and buried unsatisfied polar atoms.

4.3 Materials and Methods

4.3.1 Computational Matching and Design

Matching

The automated matching (Figure 4.2) and design stages of the grafting protocol were implemented in the Rosetta modeling platform. The matching stage exhaustively searched 13337 single chains from the PDB[4]. Loop 1 of the epitope (364 to 373, 2NY7 numbering) was defined as primary loop, and it sets the rigid body orientation between the query scaffold and the antibody. The second loop was considered secondary and honors the orientation of the initial complex. The computational algorithm takes as input the structure of the epitope loops in complex with the antibody and a file defining sub-ranges of the epitope. A C2N superposition scheme was used to search for primary loop matches, the last residue on the C-terminus of the primary epitope loop was aligned onto every residue of the query scaffolds and RMSDs were measured between the N-terminus of the loop and each residue of the query scaffold. If the RMSD between the N-terminus of the epitope and the query residue was bellow a user defined threshold (2 Å), a primary match was found. Upon identification of a primary match, the initial 2 loop epitope was recovered maintaining the original rigid body orientation, next the algorithm queried the scaffold for matches to the secondary loop. Secondary matches were searched based on the RMSD between the N and C termini residues of the second loop and two residues on the query scaffold using also a user defined RMSD threshold (5 Å). To finalize, the antibody was aligned to a hybrid-structure composed by the query scaffold with the epitope in the putative match location, and steric clashes were measured between the epitope and the scaffold (intra-clash), and between the scaffold and the antibody (inter-clash). Matches were discarded based on several criteria: the distance between the junction points of the epitope loop and the scaffold, inter-clash and intra-clash scores.

Modeling and design of the junctions between scaffolds and the epitope

The first round of design was performed using the Remodel algorithm (Huang et al. in preparation) implemented in Rosetta. The alignment between the scaffold and the epitope loops produced by the matching stage was recreated and three gaps in the polypeptide chain were closed using loop modeling and design. The starting structure for the remodel stage was obtain as follows: A) for Loop 1 the segment comprised between the residues 115-130 (PDBid:2bod numbering) was replaced by the segment of gp120 from residue 365 to 372 with residue 372 (Loop 1 gp120) being superimposed to residue 130 (2bod); B) for Loop 2 the original orientation relative to Loop 1 was recovered and residues 73 to 92 from the scaffold were replaced with residues 472 to 476 from gp120. The remodel algorithm samples backbone conformations by doing steps of fragment insertion[111] and Cyclic Coordinates Descendent (CCD)[39, 110] ensuring that the connectivity of the polypeptide chain is maintained, followed by a step of RosettaDesign[15] and full-atom minimization. The rebuilding and design steps were performed in the presence of the b12 antibody. In Remodel simulations a loop of four residues was built connecting residue 114 (PDBid:2bod numbering) to 365 (gp120 numbering) and 365 was also allowed to move to improve efficiency of the chain closure algorithm. For Loop 2 connecting regions in the N and C termini, four and three residue segments were built respectively, and both edge residues of epitope Loop 2 were allowed to move. Other lengths for the connecting segments were attempted and the reported lengths were selected based on the higher efficiency of the algorithm closing backbone gaps. The designs resulting from the remodeling stage were selected based on several scores: full atom energy, Ramachandran score and counts of polar buried unsatisfied atoms, as implemented in Rosetta++[111].

Structure-sequence diversification

The algorithm described by wang et al. and Qian et al. [39, 38] was used to generate structural ensembles of segments upstream and downstream of the epitope residues in the absence of the b12 antibody. 20,000 models for each individual segment were generated in the context of the initial design (Figure 4.3). The best 1,000 models by Rosetta energy

were selected for a following design stage and for each loop relax model 10 designs were generated; the designed positions are shown in figure 4.3. A first redundancy filter was applied to the 10000 designs to eliminate models with the same Rosetta full atom energy, reducing the number of designs by approximately 90%. Next, designs were filtered according to the Ramachandran score, the criteria used was to discard all the designs with a energy higher than the initial computational design (12 energy units). The complex between b12 and the remaining designs was recreated, followed by an interface side-chain repacking step for an accurate calculation of the interaction energy between the two molecules and as a last step of filtering all the designs with an interaction energy higher than 0 were discarded (implying that there were steric clashes at the interface). After the filtering stage, 633 designs for loop 2N, 1,202 for loop 2C and 370 for loop 1N remained.

The computational models suggested that the sampled segments would likely interact with each other. Although simulations of the 3 segments sampled simultaneously were performed, the best scoring solutions did not converge to solutions with well ordered secondary structure elements, such as the solutions observed with only one of the segments sampled (particularly for loop 2C). We acknowledged the necessity of combining the best solutions obtained for individual loops, insuring structural compatibility between them. A small set of designs selected from the single loop simulations where then explicitly recombined in an all-against-all fashion. The atomic coordinates from loop 1N variants were explicitly combined with all variants of loop 2N and resulting variants were then combined with loop 2C. Single segment designs were recombined as follows: 9 designs for loop 1N, 7 for loop 2N and 4 loop 2C; resulting in a total of 252 different combinations. The designs were selected based on Rosetta full-atom energy and a double filtering criteria where the designs on the top 20% by energy were then selected according to a packing score[112] and counts of polar buried unsatisfied atoms. After the explicit recombination of the segments one more step of design was performed in positions allowed to change in all single segment simulations. Next, 6 different designs were selected for the last step of flexible backbone sampling and design. The designed variants were selected by Rosetta energy, but also according to the diversity of the backbones. The final 6 designs incorporated 1 variant of loop 2C, 2 variants of loop 2N and 3 variants of loop 1N. According to the single segment simulations, loop 2N

potentially interacted with both other segments. Hence loop 2N was the most adequate loop to perform a final round of conformational sampling, insuring that the three segments were optimized in the presence of each other. A last step of design and filtering was performed according to the same protocol described for the single segment ensemble generation.

Library design

A small subset of designs was selected to assemble a directed library. The designs generated from the last 6 ensembles were further filtered, using a triple criteria were the designs had to be part of the top 5% by full atom energy, packing and counts of buried unsatisfied. To further reduce the number of designs after the first filter, only the top scoring designs were considered and designs with obvious flaws such as buried polar side chain atoms were discarded after visual inspection. We also considered the initial models obtained after the recombination and design step. The frequencies of amino acids for each designable position were computed and Rosetta models where different amino acids occurred were visually evaluated. To reduce the size of the library, some amino acids that occurred on Rosetta models were discarded upon visual inspection typically for one of two reasons: A) buried side chain polar atoms; B) under-packed structure. The observed frequencies in Rosetta models and amino acids included in the libraries are shown in Figure 4.11.

4.3.2 In vitro evolution and Selection

The genes encoding 2bodx_003 and 2bodx_004 designs were synthesized with optimized codon usage and subcloned into pET29b vector (Codon Devices, Genscript Corp.). For library construction, these genes were extracted and extended by PCR to add regions of homology to the pCTCONC2 vector. The random mutagenesis library was generated with the Gene Morph II Random Mutagenesis kit (Strategene) under reaction conditions that resulted in ∼4 mutations per gene. For the directed libraries, multiple oligos were used to span the desired sequence diversity for Loop 1 or Loop 2 while avoiding the addition of stop codons and limiting the number of undesired gene variants due to degenerate codons. Loop 1 or Loop 2 ultramer oligos (IDT) were transformed into double stranded DNA and

combined with two non-mutated DNA fragments extracted from the original 2bodx_004 gene (Figure 4.12) to form full-length gene variants during homologous recombination in yeast. Libraries were transformed and screened using a yeast surface display system previously described [132, 121]. Briefly, *S. Cerevisiae* EBY100 competent cells were transformed with 1μg of pCTCONC2 triple-cut vector (BamHI/SalI/NheI, NEB)and a 20x molar excess of library DNA using electroporation. Typical transformation efficiency ranged from 10^6 to 10^7 for a 1x transformation and more than 90% of the recovered sequences from the nave libraries contained full-length in-frame gene variants. The resulting yeast culture was grown at 30 ^{o}C, 250 rpm in 2% glucose C-trp-ura media and passaged at least ttwice. Once the culture reached a density of 2x10^7-3x10^7 cells/mL it was transferred to 2% galactose C-trp-ura media and grown for 14-16 hours to induce protein expression on the surface of yeast. 10^6 to 10^8 yeast cells were then pelleted and washed three times with PBSA buffer (0.01 M sodium phosphate, pH 7.4, 0.137 M sodium chloride, 1 g/L bovine serum albumin). Cells were then incubated with MAb b12 IgG on an orbital shaker for 1 hour at 4 ^{o}C. Finally, cells were pelletted, washed three times and fluorescently labeled with phycoerythrin conjugated a-hIgG (Invotrogen) and fluorescein isothiocyanate labeled a-cMyc Ab at 4 ^{o}C for 1 hour. Cells were analyzed using fluorescence activated cell sorting (BD Influx, BD Biosciences) and double positive clones were sorted, expanded, induced and labeled as before for additional rounds of screening. Decreasing MAb b12 IgG concentrations were used for labeling subsequent selection rounds to reduce library diversity and select high affinity clones. Individual clones were selected after each round and their DNA was isolated using the Zymoprep II kit (Zymo Research). Standard PCR was used to amplify the gene of interest from this DNA and the resulting product was sent for sequence analysis.

4.3.3 Protein Expression and Purification

DNA segments encoding scaffold constructs were synthesized with optimized codon usage and RNA structure (Codon Devices, Genscript Corp.), subcloned into pET29 (EMD Biosciences) and transformed into Arctic Express^{TM}*E. coli* (Invitrogen). Single colonies from the transformation were grown overnight at 37 ^{o}C in starter cultures with Luria Broth (LB)

plus Kanamycin (100 mg/ml). 10 mL of starter cultures were expanded into 1 L of LB plus Kanamycin (100 mg/mL) and incubated at 37 ^{o}C, 250 μM of IPTG were added to the culture when log phase was reached and incubated overnight to induce protein expression at 12 ^{o}C. Cultures were then pelleted and resuspended in start buffer (160 mM Imidazole, 4 M Sodium Chloride, 160 mM Sodium Phosphate), a tablet of protease inhibitor (Roche) was added and the cell suspension was frozen at -20 ^{o}C. The cell suspension was thawed and 10ml of 10x Bugbuster (Novagen), 50 μL of Benzonase Nucleases (Novagen) and 1.7 μL of rLysozyme (Novagen) were added to lyse the cells, the suspension was then gently tumbled in an orbital shaker for 20 minutes. Lysed cells were pelleted and the supernatant was filtered through a 0.2 micron filter (Millipore). Supernatants were tumbled with 5 mL of nickel-NTA resin (Fisher) for 1 hour at 4 ^{o}C. The resin was washed then rinsed with Wash Buffer (50 mM imidazole, 500 mM Sodium Chloride and 160 mM Sodium Phosphate) and eluted with elution buffer(250 mM Imidazole, 500 mM Sodium Chloride and 20 mM Sodium Phosphate). Fractions containing the construct of interest were combined and further purified by preparative size exclusion chromatography (SEC) on Superdex 75 16/60 (GE Healthcare) at room temperature in PBS or HBS. Fractions containing the protein of interest were collected, confirmed by SDS page and concentrated by ultrafiltration (Vivaspin, Bioexpress). Protein concentration was determined by measuring UV absorption signal at 280 nm and calculated from known amino acid composition extinction coefficient. Site-directed mutagenesis was performed following the "Kunkel" method[133, 134].

4.3.4 Solution Oligomerization State and Circular Dichroism

The monodispersivity and molecular weight were further assess by analytical SEC (Agilent, 1200 series) coupled to an on-line static light scatter (SLS)(miniDAWN TREOS, Wyatt). The molecular weight determination was performed with the ASTRA software (Wyatt). Solution thermostabilities (T_m) were determined by circular dichroism with an Aviv 62A DS spectrometer. Far-UV wavelength scans (260 nm - 190 nm) at approximately 20 μM protein concentration were collected in a 1 mm path length cuvette. Temperature-induced protein denaturation was followed by change in ellipticity at 220 nm.

4.3.5 Biacore

All experiments were carried out on BIACORE 2000 at 25 ^{o}C. Running buffer HBSEP (0.01 M HEPES pH 7.4, 0.15 M NaCl, 3 mM EDTA and 0.005% (v/v) Surfactant P20) (GE Healthcare). For analysis the human antibodies were captured on Research grade CM5 sensor chip containing 8000-9000 RUs of Anti-Human IgG from Human Antibody Capture kit (GE Healthcare). The Anti-Human IgG was immobilized in all four flow cells including a reference flow cell at similar levels using Amine Coupling kit (GE Healthcare) according to the manufacturers manual. All data were collected using two replicate injections for each concentration of analyte. The regeneration was performed when needed with 60 seconds injection of 3 M $MgCl_2$ from Human Antibody Capture kit. Flow rates during the experiment were maintained at 50-100 μl/min. Data preparation and analysis were performed using Scrubber software (version 2.0, BioLogic Software,Campbell, Australia). For kinetic analysis biosensor data were globally fit to a mass transport limited simple bimolecular binding model:

$$A_0 \underset{k_m}{\overset{k_m}{\rightleftharpoons}} A + B \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} AB$$

A_0 represents injected analyte

For equilibrium analysis each data set was fitted to a single site interaction model:

$$R_{eq} = \frac{R_{\max} * C_A}{C_A + K_D}$$

R_{eq} - response value at equilibrium

C_A - concentration of the analyte

R_{max} - maximum response when all binding sites are occupied by the analyte

K_D - dissociation constant

4.4 Results

4.4.1 Computational Design

The Protein Data Bank was searched for acceptor scaffolds able to accommodate the b12-bound conformation of the two main epitope-loops from gp120. A detailed description of the matching algorithm is given in the methods section and also outlined in Figure 2. The matching strategy used to select the discussed scaffold (2bodx) was C2N match in the primary loop (Loop 1) and end point match for the secondary loop (Loop 2). In a C2N match the residue on the C terminus of the loop is superimposed on each residue of the query scaffold, and RMSDs of the N terminus residue were measured to all the residues of the query protein. Alignments where the RMSD was lower than a user-defined threshold (2 Å) were kept, resulting in a primary match. Next, Loop 2 was aligned in all the primary matches to recreate the orientation of the b12-gp120 complex. RMSDs were measured between both ends of Loop 2 and residues in the query scaffold, and if the values were bellow a user-defined threshold (5 Å) the alignment of the two loops on the scaffold was accepted as a match. Matches were then filtered by backbone clashes between epitope and scaffold (defined as intra-clash), and backbone clashes between the scaffold and the antibody (defined as inter-clash). Several variants of the matching schemes were attempted: Loop 1 superposition with Loop 2 end point match; Loop 2 superposition with Loop 1 end point match. Multiple variants based on 11 different scaffolds were experimentally characterized. Most of the proteins were barely soluble or insoluble with the exception of the ones based on 2 different scaffolds one of them the 2bodx scaffold.

The choice to focus in 2bodx scaffold was made primarily for three reasons: the matching scheme allowed to take advantage of the only ordered secondary structure element in the target epitope a helical turn in the basis of Loop 1 (C terminus), which served as an "anchor" for the placement of the epitope on a helix terminus; upon a search for homologues of 2bodx the region remodeled showed high structural plasticity (Figure 4.9), strongly suggesting that the scaffold region could tolerate multiple sequence/structure solutions; the experimental feedback from the first round, where most of the 2bodx designs were expressed in *E. coli* and soluble to high concentrations, and one of the designs also bound weakly to b12 but

specifically by SPR (Figure 4.10).

From the first round of pure computational design, 8 constructs were expressed in *E. coli*, 6 were soluble and purifiable using immobilized metal affinity chromatography and SEC. The solution oligomerization state was analyzed by Static Light Scattering (SLS) and SEC and the designs were monodisperse and monomeric in solution. SPR was performed to assess binding to antibody, 2 of the designs showed signal above background (2bodx_003 and 2bodx_004), with 003 showing a higher response than 004. The difference between the two constructs was that 003 had a disulfide bond further in space and in sequence from the modeled region, but that could potentially impact b12 binding according to the computational model. The extremely low affinity assessed by SPR made the specificity of the scaffold-b12 interaction questionable. To clearly address the binding specificity issue, a knockout variant (D140R) of the 003 design was tested by SPR, and binding was abolished by this single point mutation that has been reported to have the same effect in the b12-gp120 interaction[135].

The results for the initial computational design showed two desirable features for the following optimization stage: the scaffold tolerated a large number of mutations including deletions and insertions, while remaining a soluble and well behaved molecule; and specific binding to the b12 antibody, with approximately 10,000 fold lower affinity than that of gp120.

4.4.2 Computationally Inspired Libraries and In Vitro Evolution

Following the results from the computational design round, we undertook a mixed computational and *in vitro* evolution approach to generate and screen multiple variants of 2bodx_003 epitope-scaffold. We hypothesized that optimizing the connecting regions between the epitope loops and the scaffold would lead to better epitope stabilization, resulting in higher affinity for b12 MAb. Given the large number of residues involved in epitope stabilization (16) (Figure 4.3) as well as the possible interactions between the two loops, an effective way to investigate this large sequence/structure space was to first computationally determine energetically favorable sequences for Loop 1 and Loop 2 connecting segments. From the

best models, the resulting amino acid diversities at each of the 16 critical positions were merged and analyzed as two directed libraries displayed on the surface of yeast.

The computational protocol of sequence-structure diversification was devised on a philosophy of "divide and conquer". First, the connecting segments of Loop 1 N-terminus, Loop 2 C-terminus and Loop 2 N-terminus were individually modeled to ensure fine sampling of the structural segments (Figure 4.3). The sequence-structure diversification was achieved using Rosetta loop prediction[39] for the generation of structural ensembles and RosettaDesign[15] for sequence diversification. The resulting models were evaluated based on several of the implemented Rosetta energies, such as the Ramachandran distribution of the backbone dihedral angles[111], packing[112] and buried unsatisfied polar atoms counts. In a second step, the best solutions for the individual segments were combined iteratively and additional computational modeling ensured their structural compatibility (Figure 4.3).

The resulting models have better Rosetta metrics than the original 2bodx_003 design, suggesting that more energetically favorable regions of the sequence/structure landscape were indeed sampled. This iterative procedure might thus serve as an efficient way to avoid entrapment in local energy minima, similarly to previously described strategies[48].The best 45 models were aligned and showed sequence diversity at positions 72-76, 80-83, 86 and 107-110, 121, 125 for Loop 2 and Loop 1 connecting regions respectively (Figure 4.4). Residue frequencies at these positions (Figure 4.11) showed good sequence convergence in some cases (76, 86, 110), but most often no consensus residue was identified. Furthermore, at some positions (72,107,109) there was no preference for a given type of amino acid (i.e. polar, hydrophobic), underscoring the vast structure and sequence space explored during the modeling as well as the large number of conformations of the loop connecting regions.

To investigate the diversity present in the best 45 models, two directed libraries were built and screened for b12 MAb binding by florescent activated cell sorting on the surface of yeast[132, 121]. Although it would have been ideal to investigate both Loop 1 and Loop 2 diversity simultaneously, the large number of possible combinations (10^{12}) compared to the regular size of yeast libraries (10^{7}) made this impractical. To further reduce the library size, the best 45 models were visually inspected and some residue identities were excluded based on their low frequency or other undesired characteristics such as buried side chain

polar atoms.

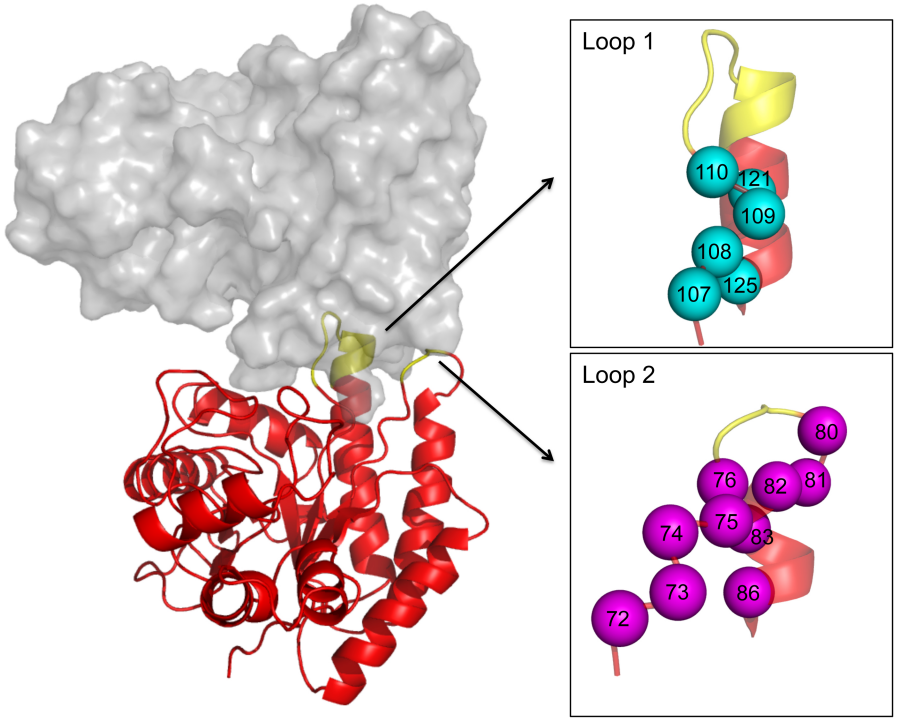


Figure 4.4: Residues sampled in computational design calculations to inspire directed libraries. The C-alpha atoms are shown in spheres being labeled with the residue number.

The first library (Library 1) (Figure 4.11) covered the sequence diversity of Loop 2 ($4x10^6$) combined with 8 variants of Loop 1 present in 23 out of the 45 models that were used to generate the library. Including multiple Loop 1 variants was deemed important to avoid any potential incompatibility between Loop 1 and Loop 2 sequences. The second library (Library 2) (Figure 4.11), covered the sequence diversity of Loop 1 ($2x10^5$) combined with the best Loop 2 variants identified from Library 1. Additional diversity was added to the libraries due to codon degeneracy, and some of the previously modified residues in 2bodx_003 were allowed to revert to their native identity since they were considered non-essential for epitope stability. Overall, 21 positions were allowed to change with different degrees of variability in the two libraries.

Before screening Libraries 1 and 2 , a random mutagenesis library was constructed to

optimize the yeast surface display levels of 2bodx_003. This resulted in the isolation of clone 2bodx_Y3 which was different by two mutations, A118V and S177G, compared to the computational design. S177G removed a putative glycosilation site which might have interfered with protein folding in yeast, while A118V added a contact present in the gp120-b12 complex to 2bodx_003. 2bodx_Y3 not only had increased cell surface display levels, but also showed higher affinity for b12 (Table 4.1, Figure 4.6). The S177G mutation was included in Libraries 1 and 2, but A118V was not added at this stage since the computational modeling did not allowed any variability at that position.

The Libraries were generated and transformed by taking advantage of yeast homologous recombination. Specifically, Loop 1 and Loop 2 encoding DNA fragments were mixed with non-variable regions of the 2bodx_003 gene to form full length genes when transformed with pCTCONC2 vector (Figure 4.12). This approach facilitates library construction and adds an additional diversification step due to DNA shuffling[121].

Library 1 was screened with decreasing b12 IgG concentrations (1 μM, 1 μM, 100 nM, 10 nM, 1 nM) for 5 rounds (Figure 4.5 and Figure 4.13). Increased b12 binding was observed during the selections, and after the fourth round the library complexity was reduced to 3 different clones. An additional round of selection resulted in convergence to clone 2bodx_032. To construct Library 2, we decided to combine all the variants present after the third selection round of the Library 1 with the computationally determined diversity of Loop 1 (Figure 4.11). The DNA of 48 individual clones from the third selection round of Library 1 was sequenced and 18 different sequences were identified, with 12 of them represented more than once. We choose to include several Loop 2 variants for Library 2, rather than just the single converged clone (2bodx_032), to account for the possibility that in the presence of different Loop 1s other Loop 2s might be better for b12 binding due to their structural proximity. Indeed, after 3 rounds of selection at 10 nM, 100 pM and 10 pM b12 IgG respectively, Library 2 converged to a single clone, 2bodx_042, which had a sequence for Loop 2 different from 2bodx_032. This Loop 2 sequence was present in only 2% of the clones from the third selection round of Library 1. A third library (Library 3) was devised to optimize buried or partially buried residues underneath the epitope loops that were under-sampled in Libraries 1 and 2.

Screening Library 1 with b12 IgG was advantageous because it allowed identification of low affinity binders due to avidity effects. However, Library 2 showed increased surface display levels compared to Library 1 (data not shown), raising the possibility that clones were being selected for increased avidity instead of affinity. To eliminate this possibility, Library 2 was re-screened with biotin-b12 Fab (data not shown) and it converged to the same 2bodx_042 clone as when screened with b12 IgG.

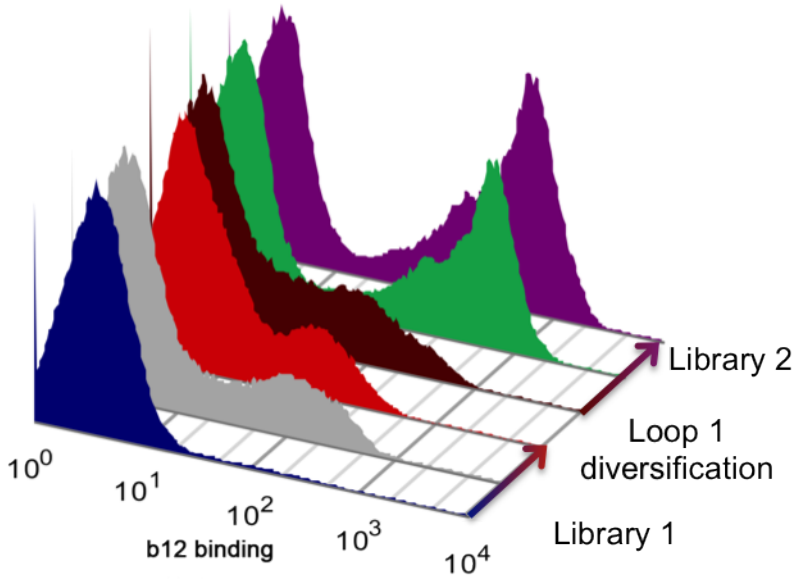


Figure 4.5: Sequential screening of directed libraries to obtain an improved b12 binder. A library targeting Loop 2 (Library 1) connecting segments was screened in the presence of 8 variants of Loop 1. Sequence convergence for Library 1 was reached after 5 rounds of selection, resulting in 3 different clones. A library targeting Loop 1 (Library 2) connecting segments was screened in the presence of 18 Loop 2 variants selected from the 3^{rd} selection round of the Library 1.

4.4.3 Affinity and Specificity

The 2bodx variants identified from the selections experiments were expressed recombinantly in *E. coli* and biophysically characterized. Dissociation constants (K_Ds) for the b12 interaction of multiple clones were determined by SPR (Table 4.1, Figure 4.6).

2bodx_Y3 had a K_D for b12 of 30 μM, a 10 fold improvement over 2bodx_003, most likely due to additional contacts made by V118 with the Ab. This residue is also present at the corresponding position on gp120 and interacts with b12 in the crystal structure of the complex. However, V118 cannot be modeled on 2bodx_003 without introducing a steric clash between the epitope-scaffold and the CDRH3 of b12. This suggests inaccuracies either in the scaffold-antibody rigid body orientation or unaccounted flexibility from the CDRH3 of b12 which may alleviate the clash. 2bodx_032 had the same Loop 1 sequence as 2bodx_003, but differed by 8 mutations in loop 2, which resulted in a b12 affinity of 13 μM. The highest affinity clone isolated from Library 2 (2bodx_042) bound b12 with a K_D of 166 nM. This variant had 17 mutations relative to 2bodx_003 distributed over both the Loop 1 and Loop 2 connecting regions. Introducing the A118V mutation on 2bodx_042 resulted in a new variant (2bodx_043) which bound b12 with a 33 nM affinity. Two additional mutations (P83 and M124) selected in Library 3 improved b12 affinity to 10 nM. Sequence changes for all the mentioned clones can be found in Table 4.2.

Table 4.1: Kinetics and equilibrium b12 binding constants for selected clones were assessed by SPR. b12 binding affinity of the 2bodx_43 clone is in the same range as the b12-gp120 interaction.

Variant	**Method**	**b12 binding affinity and kinetics by SPR**			
		K_{on} ($M^{-1}s^{-1}$)	K_{off} (s^{-1})	K_D (kinetic)	K_D (equilibrium)
003	initial design	-	-	-	>300 μM
Y3	rand. library(RL)	-	-	-	≈30 μM
32	Library 1(L1)	-	-	-	13 μM
42	L1+ Library 2(L2)	$1.35x10^6$	0.234(6)	173(6) nM	166.6(5) nM
43	L1+L2+RL	$3.09x10^6$	0.107(2)	34.6(7) nM	33.5(1) nM
45	L1+L2+RL+L3	$3.82x10^6$	0.039(3)	10.33(2) nM	10.3(1) nM

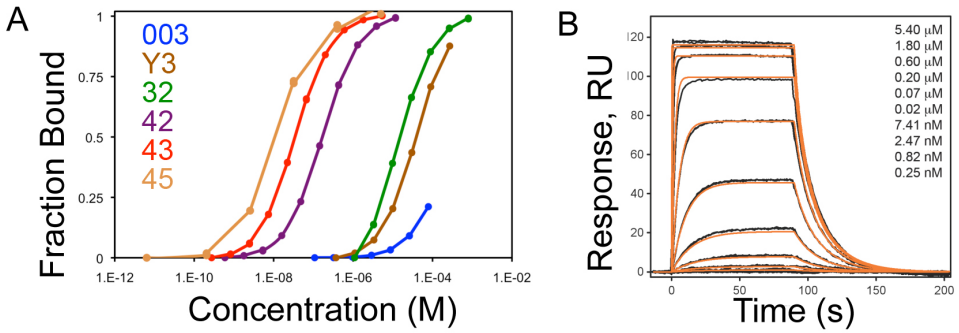


Figure 4.6: Equilibrium curves and sensorgrams assessed by SPR. b12 binding affinity of the 2bodx_43 and 45 clones are in the same range as the b12-gp120 interaction. (A) Equilibrium binding curves determined by SPR. (B) SPR kinetic analysis of 2bodx_43 interacting with b12 IgG. The kinetic fits are shown in red.

The affinity range of the best clones is similar to the K_D of b12 for gp120 (20 nM [74]) and 30,000 times better than the original computational design. To probe the specificity of 2bodx_043, residue D104 was changed to R, which resulted in complete loss of binding to b12 (Figure 4.8), similar to the effect of the D368R mutation on gp120. Circular Dichroism experiments showed that the protein was stable ($T_m \sim 75\ ^oC$) and well folded (Figure 4.7A and B). Static light scattering analysis revealed that 2bodx_043 was monomeric (Figure 4.7C).

The CD4 binding site (CD4bs) in gp120 is a known site of interaction for multiple antibodies against HIV-1. To evaluate the specificity of the transplanted binding site in the context of the scaffold,the binding of 2bodx_043 to CD4 and a panel of 9 CD4bs antibodies was assessed by SPR. 2bodx_43 was highly specific for b12 MAb and did not interact with any other CD4bs Abs or with CD4 itself even at high concentrations (Figure 4.8). Based on the structural information available for the antibodies tested, the lack of binding is due to either insufficient contacts made with 2bodx_043 (CD4, VRC01, etc.), to the different conformations these Abs induce in Loop 1 and 2 upon binding, or to steric barriers present on the scaffold but absent from gp120. The b12 specificity shows that through our combined *in silico/in vitro* approach we successfully transplanted a complex binding site known for its promiscuity, and stabilized it in the precise conformation to recognize only its target ligand

Table 4.2: Sequence changes of the selected clones relative to the initial computational design. Residue numbers are indicated on the top of each sequence column. The epitope residues are colored in red.

Library	Selection round	Variant	Loop 2												Loop 1														
			72	73	74	75	76		80	81	82	83	84	86	107	108	109	110		118	119	120	121	122	124	125			
		CD	N	P	D	P	G	GDM	D	N	G	F	E	G	D	P	G	G	SGGDPEI	A	E	A	A	W	F	A			
RL	-	Y3	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.		V	.	.	.	.	.	.			
L1	Sort 5	32	E	.	Y	.	T		G	K	.	W	A	A	.	.	.	.			.	.	.	.	.	.			
L2	Sort 3	42	D	V	G	.	T		S	Q	.	E	A	.	Y	.	L	Y		.	Q	E	W	L	T	V			
RL+L2	-	43	D	V	G	.	T		S	Q	.	E	A	.	Y	.	L	Y		V	Q	E	W	L	T	V			
L3	Sort 3	45	D	V	G	.	T		S	Q	.	P	A	.	Y	.	L	Y		V	Q	E	W	L	M	V			

CD - Computational Design

RL - Randoom Library

L1- Library 1

L2 - Library 2

L3 - Library 3

with high affinity.

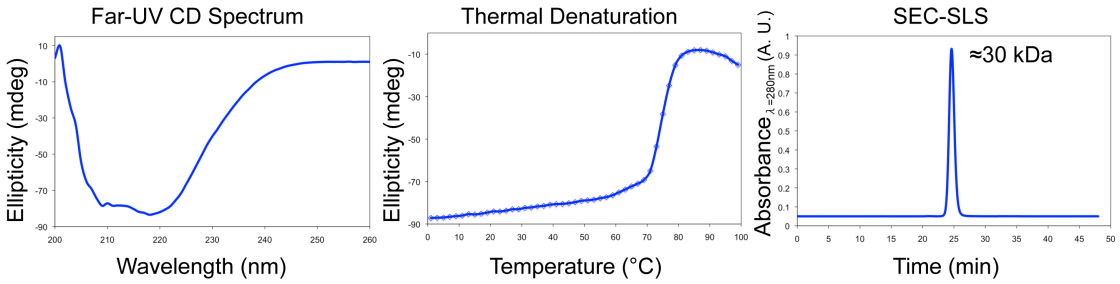


Figure 4.7: Thermal stability and solution oligomerization state. 2bodx_43 is folded and monomeric. Circular Dichroism shows a characteristic spectrum of a folded α/β protein. Thermal denaturation shows cooperative unfolding with a T_m of ~75 ^{o}C. Size exclusion cromatography (SEC) and static light scattering (SLS) show a monomer in solution with the expected molecular weight of ~30 kDa.

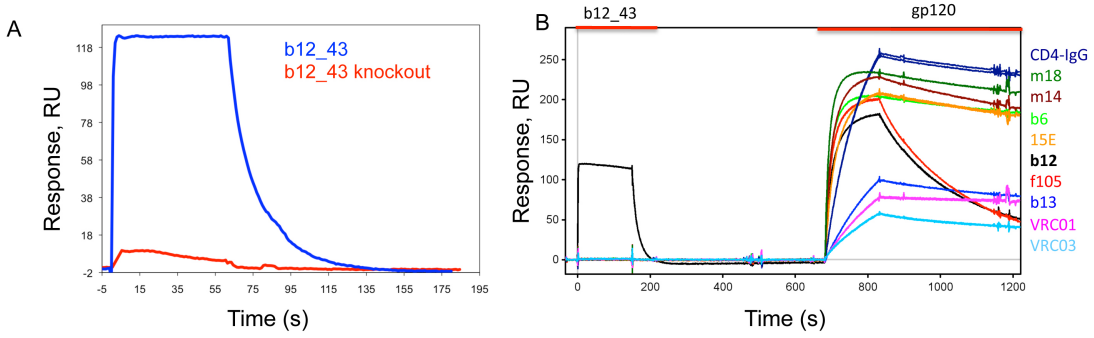


Figure 4.8: 2bodx_43 shows highly specific binding to b12. (A) Knockout mutation from the b12-gp120 interaction shows the same effect in 2bodx_43. (B) Multiple antibodies that bind to the CD4 binding site of gp120, 2bodx_43 only binds to b12.

4.5 Discussion

To our knowledge, we accomplished the first complex epitope transplantation to a heterologous protein using computational methods together with *in vitro* evolution. The aggressiveness inherent to this design problem is considerable, where the scaffold was "sculpted" to accommodate a complex epitope, including changes on the length of several segments as well as drastic alterations in their secondary structure. The devised strategy enabled the design of a starting sequence with weak binding activity (K_D ~300 μM) that was further optimized using *in vitro* evolution leveraged by computationally inspired libraries to achieve an improvement in affinity on the order of 30,000 fold. The best epitope-scaffolds bind b12 with dissociation constants in the same range of the original b12-gp120 interaction (20 nM). The epitope-scaffold is highly specific, engaging the predefined target antibody but not other ligands known to engage the same site on gp120, further validating the rational strategy taken.

The computational methodologies presented are applicable to a wide range of challenges dealing with backbone grafting, spanning the single segment to multiple segments grafting (2 or more). The matching scheme is able to perform multiple types of alignments, allowing the selection of the most appropriate scaffolds according to intrinsic features of the target binding site. Given the aggressive remodeling required, a flexible backbone design approach was devised to optimize individual remodeled segments that were then recombined to generate a directed library. The library was efficiently and rapidly screened in a yeast display format. Given the observed improvements, it is appealing to argue that routine use of flexibility in protein design will enable greater steps in the improvement of desired functions.

The results described open several novel routes to the field of vaccine design, where strategies to generate candidate molecules have been limited to variations based on the pathogen's native molecules, or scaffolding of linear epitopes by side chain grafting (Chapter 2). The specificity shown by the scaffold is a key feature for promising immunogens, since the main aim is to re-elicit a particular kind of neutralizing antibodies and dampen the elicitation of all the non-neutralizing ones. The possible applications are not limited to vac-

cine design but also for other applications that require extremely aggressive computational design such as enzyme design or protein interaction inhibitors.

In order to benchmark properly the computational design protocol, crystal structures are being solved. Similarly to evaluate the immunogenic profile of the designs, immunization experiments are ongoing.

4.6 Acknowledgements

Bruno Emanuel Correia (B.E.C.) performed computational design calculations. Mihai Azoitei (M. A.) and B.E.C designed directed libraries. M. A. performed *in vitro* evolution and biophysical characterization. Oleksander Kalyuzhniy performed biophysical characterization. Chris Carrico and Roland Strong performed crystallization. Yih-En Ban wrote the matching code. Alex Schroeter performed protein purification. William Ray Schief (W.R.S.) supervised the project. B.E.C, M. A. and W. R. S. wrote the manuscript.

4.7 Supplemental Material

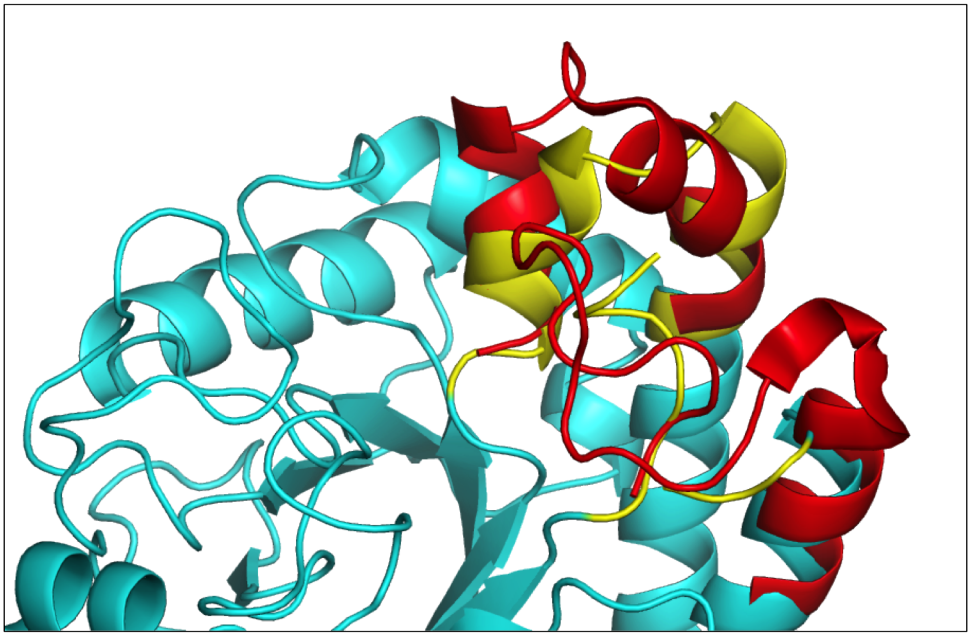


Figure 4.9: Structural diversity of 2bodx homologues. Corresponding regions to the remodeled segments of 2bodx homologous show an appreciable structural diversity. 2bodx is colored in cyan and the remodeled regions were colored in yellow. For clarity only the corresponding segments to the 2bodx remodeled regions of one homologue (PDBid: 1gz1) are shown in red. The observed diversity suggested that these particular segments could be designed without impacting the overall structural stability.

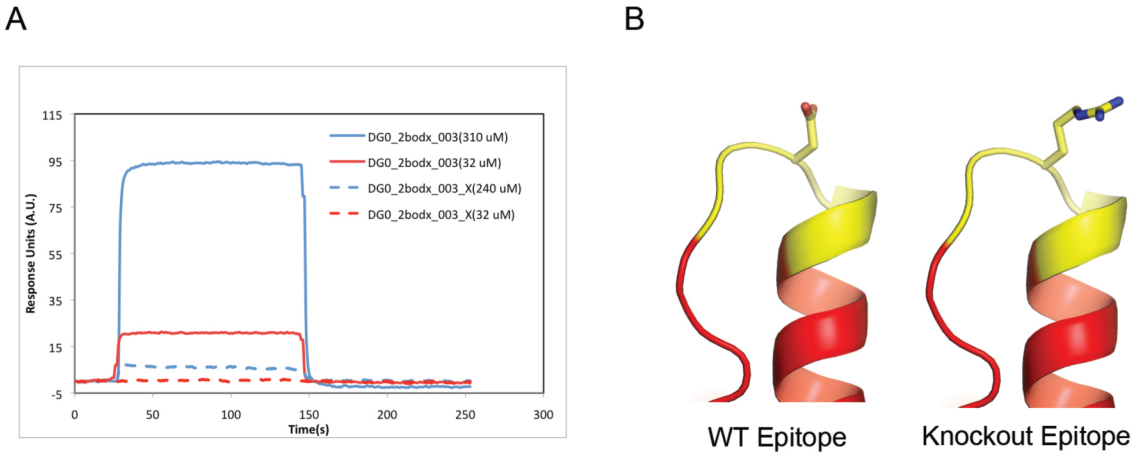


Figure 4.10: B12 binding of a knockout mutant of the initial computational design. (A) SPR sensorgrams where 2bodx_003 (solid lines) shows higher responses at equivalent concentrations comparing to its knockout version (dashed lines). (B) The knockout mutation in gp120 that depletes b12 binding is D368R. A model of the equivalent mutation in 2bodx is shown.

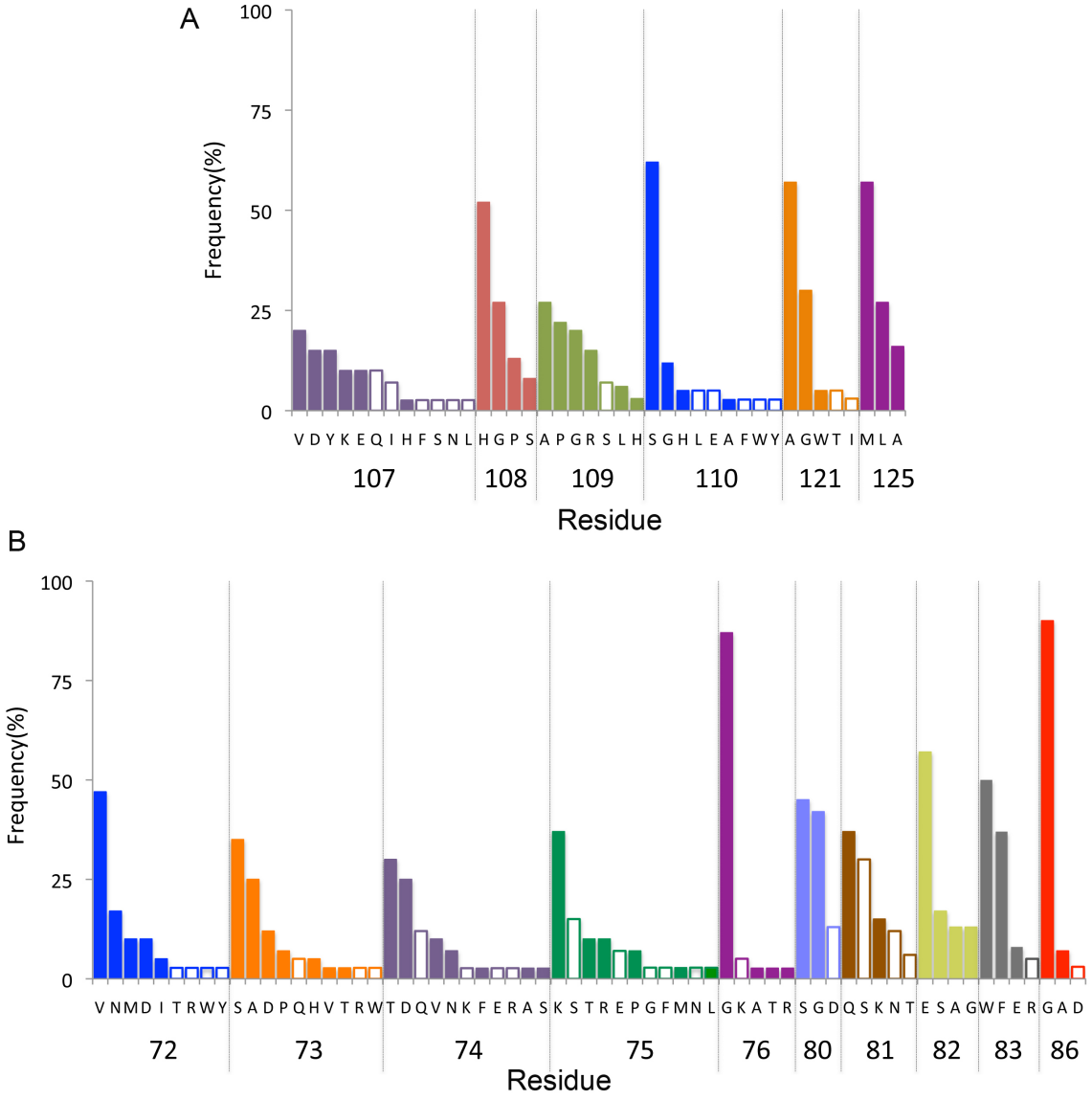


Figure 4.11: Amino acid frequencies in residues designed with Rosetta. After a selection of the designed models according to structural quality indicators, amino acid frequencies were computed for each residue. Filled bars indicate the amino acids included in the directed library. (A) Positions considered for Loop 2 region. (B) Positions considered for Loop 1 region.

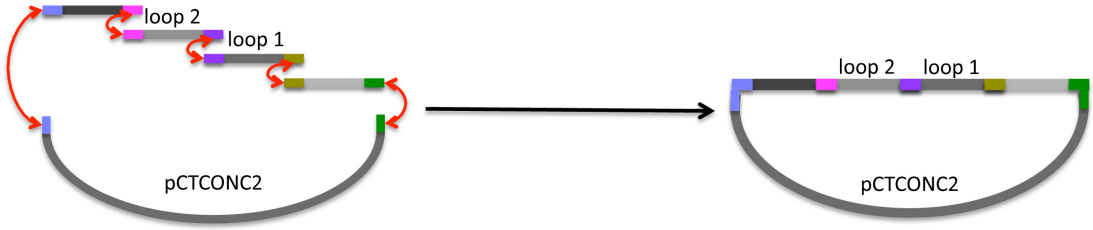


Figure 4.12: Homologous recombination strategy to screen sequentially directed libraries. The 2bodx gene was split into four segments by PCR, 2 of them had no positions targeted by the library and another two that encoded Loop 1 and Loop 2 of the b12 epitope. Initially the full Library 1 (Loop 2) was screened in the presence of eight variants of Loop 1 and the Library 2 (Loop 1) was transformed together with sequences of an early round of selection with at least 18 different sequences for the Loop 2 region.

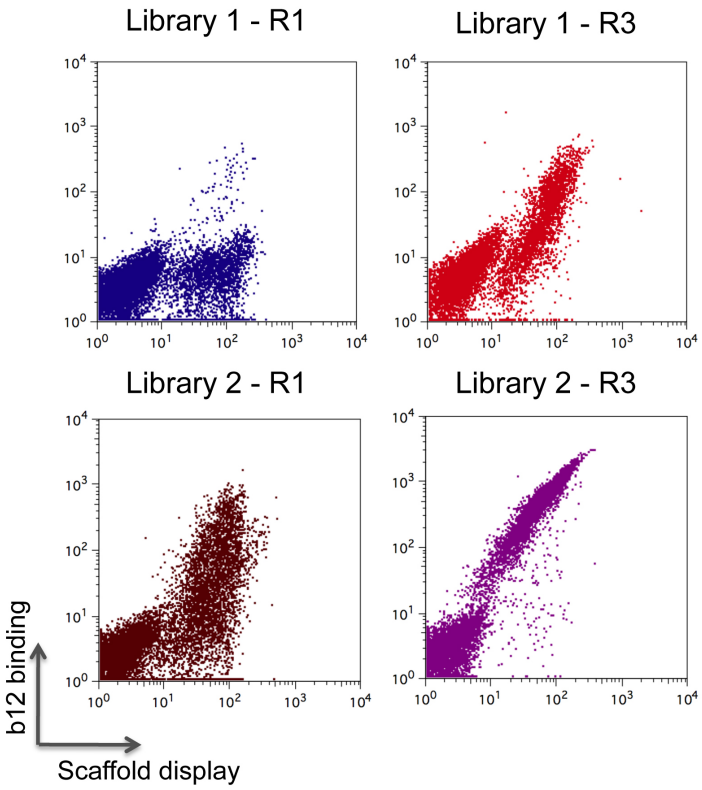


Figure 4.13: Facsograms of the 1st and 3rd rounds of sorting for Libraries 1 and 2. The sorts of Library 1 were performed with 1 μM and 100 nM b12 for the 1^{st} and 3^{rd} round respectively. For Library 2 the concentrations used were 10 nM and 10 pM in the 1^{st} and 3^{rd} sort.

Chapter 5

FOLD FROM LOOPS - FOLDING PROTEINS TO STABILIZE FUNCTIONAL SITES

5.1 Abstract

Many advances in computational protein design have been accomplished in the past few years. Despite all the progress, encoding function into computationally designed proteins remains a difficult task. Typically, to engineer function into proteins a small number of residues, predicted to confer the desired functionality, are designed on a naturally occurring protein scaffold able to accommodate these mutations. Functions such as the design of catalytic activity, creation of a binding sites and specific small ligand receptors have been successfully engineered. These methods to engineer function usually treat the scaffold as a rigid body, allowing little or no backbone flexibility. While this strategy has been successful, there are several bottlenecks for the routine design of function into proteins. The potential scaffold space is limited to folds observed in nature for which a high-resolution crystal structure has been solved. Additionally, the lack of backbone flexibility in the current approaches further limits the usage of the available scaffolds and hinders the ability to identify potential neighboring residues that can help coordinate the functional side chains. Here we describe a novel computational protocol called Fold from Loops (FFL) that functions quite differently to other current strategies. Rather than starting with a protein of known structure and mutating a small set of side chains to achieve functionality, FFL starts with a structural motif such as a binding site that already has functionality, and builds a protein around it. The structural motif is treated as a folding nucleus being held fixed and the target topology is folded around the structural motif, loosely guided by C-alpha constraints providing maximal support and stabilization to the folding nucleus. This protocol was used to stabilize an epitope of the Respiratory Syncytial Virus (RSV) recognized by the broadly neutralizing (bn) antibody (Ab) Motavizumab. Several of the designed proteins expressed, were soluble,

thermally stable and some showed pM affinities to the Motavizumab Ab, in the same range of affinities measured between the RSV F protein and the bNAb. The results suggest that the epitope was transplanted with high accuracy and the desired molecular mimicry was achieved. The computational protocol is general and readily applicable to other design problems, and we hypothesize that the degree of flexibility allowed in this protocol contributed critically for the high affinity. It will be challenging to apply FFL to design problems with discontinuous structural motifs and to fold topologies not yet observed in nature imposing functional constraints.

5.2 Introduction

The ability to rationally design proteins tailored to perform biological functions *a la carte* is a long-standing challenge of computational protein design methods. Notable milestones have been achieved in the recent past, as the design of enzymes for reactions that lack natural catalysts[31, 30, 32], modulation of the activity of transcriptions factors[136], newly reengineered specificities of endonucleases[28] and several successes on manipulating binding affinities and specificities[26, 24, 102, 103]. Many of the protein interactions observed in nature have poor binding affinities due to the need of transient and promiscuous interactions in complex biological systems[137, 138]. Nevertheless, the design of high affinity interactions achieved either by *de novo* interfaces or by "polishing" structurally characterized interaction motifs has many potential applications ranging from the basic dissection of cellular pathways to biomedical applications, such as the design of small protein inhibitors or carrier scaffolds transiently exposed epitopes in viral proteins. Thus far computational design has been mostly restricted to the "inverse folding" approach[2, 5] where the design objects have limited or no flexibility at all. Nevertheless, a basal level of flexibility is always present in proteins and there has been active development in conformational sampling techniques for design and prediction. The major hurdle in flexible backbone design is to sample realistic conformations. Remarkable successes have been reported such as the design of a new fold[21], the design of loop with sub-angstrom accuracy[48] and specificity switch using flexible loop backbone[49]. The first two examples aim for a purely structural validation with no functional constraints, the last example has a functional purpose but most

of the protein that supports the functional loop is structurally rigid. As eloquently stated by Havranek et al.[139], design of functional protein requires satisfaction of functional constraints and simultaneous energetic optimization, we perceived the need for an approach orthogonal and complementary to previous reports. Our strategy to satisfy both functional and energetic constraints was to select a functional site of interest that *in silico* acts as the folding nucleus and to fold a target topology around the functional site, in such a way that the atomistic detail of the folding nucleus is maximally supported by the surrounding polypeptide chain (Figure 5.1).

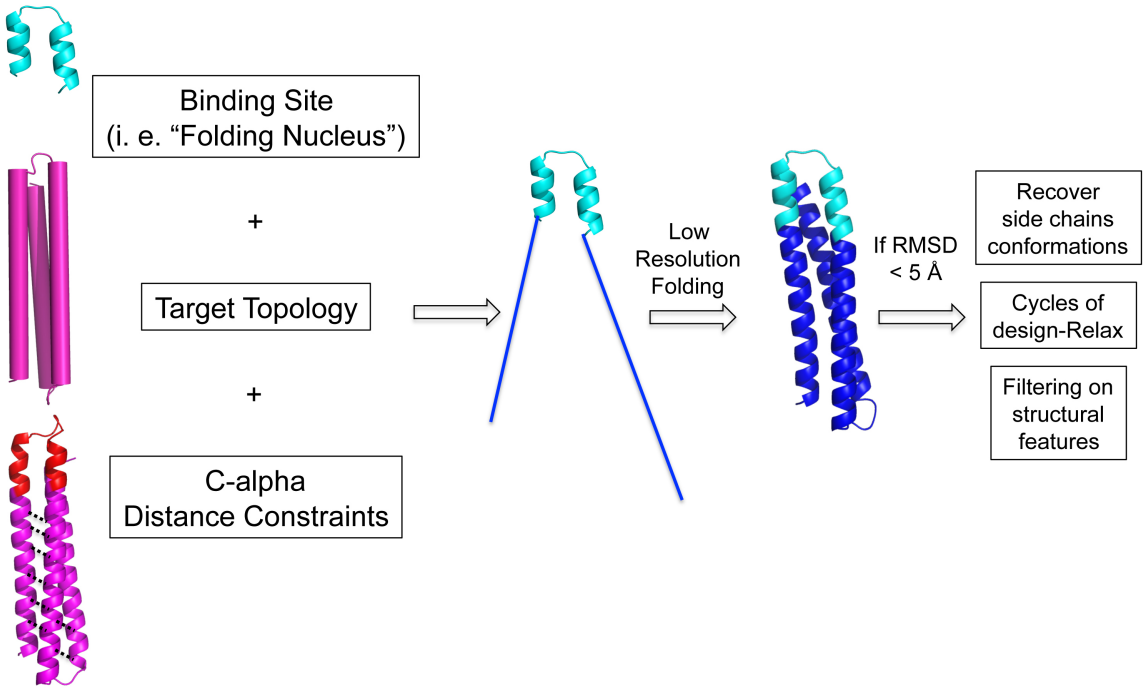


Figure 5.1: Overview of the computational procedure - Fold From Loops. The procedure takes a functional site (such as the helical hairpin represented in cyan) that will be used as the folding nucleus and remain fixed throughout the procedure. A topology has to be supplied (magenta) and distance constraints are derived from the input structure to guide the folding trajectory. The segment in red highlights the location of the folding nucleus. The polypeptide chain is extended (blue) from the folding nucleus and the chain in then folded. If the models produced are less than 5 Å away from the starting topology they are discarded, if not are taken through cycles of design and full-atom optimization.

The procedure takes a functional site (such as the helical hairpin represented in cyan

- Figure 5.1) that will be used as the folding nucleus and remain fixed throughout the procedure. A topology has to be supplied (magenta - Figure 5.1) and distance constraints are derived from the input structure to guide the folding trajectory. The segment in red highlights the location of the folding nucleus. The polypeptide chain is extended (blue - Figure 5.1) from the folding nucleus and the chain in then folded. If the models produced are more than 5 Å away from the starting topology, they are discarded. Otherwise, they enter cycles of design and full-atom optimization.

The selected functional site was a conserved epitope on the Fusion Glycoprotein (F protein) from the Respiratory Syncytial Virus (RSV). RSV is a serious public health problem, particularly to infants and children, causing 160,000 deaths per year worldwide[140]. The only available treatment is through passive immunization using the humanized monoclonal antibody Palivizumab[141]. Due to the potency and breadth of neutralization, this epitope is an important vaccine target. The epitope is the binding site through which the bn Ab Palivizumab and an affinity-matured version of Palivizumab, Motavizumab, neutralize the virus. There is no co-crystal structure of Palivizumab Ab with its epitope, but the peptide epitope is shown to be a helical hairpin when co-crystallized with the Motavizumab Ab (Figure 5.2). Furthermore, while there is no crystal structure of the F protein, there are some indications that the epitope might be only transiently exposed[140], making the F protein an undesirable immunogen. We sought to transplant and stabilize the Motavizumab epitope as found in the crystal structure with Motavizumab for use as an immunogen to re-elicit Motavizumab-like antibodies.

The target topology selected was a three-helix bundle, topologically compatible with the structure of the functional site, which is a helical hairpin[140] (Figure 5.2), nevertheless the structure of native helical hairpin in the bundle has a different residue length and a structural alignment reveals an RMSD of 2.6 Å between the backbone of the helical segments (Figure 5.2). Helical bundles have been extensively designed[142, 143, 144], several rational strategies and multiple functional groups have been incorporated in the helical bundles[145]. Several groups used existing coiled coil structures as static structural templates and imposed a known sequence pattern observed in coiled coils, the heptad repeat[146]. This new computational method used to stabilize the RSV epitope allows sampling of greater backbone

flexibility and sequence diversity, while still constraining the structure to accommodate the functional motif.

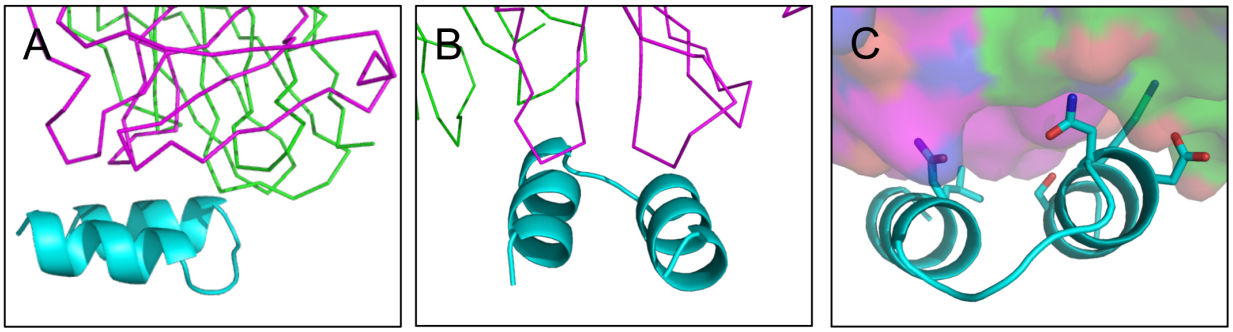


Figure 5.2: Motavizumab in complex with peptide-epitope from F-protein of RSV. (A) Side-view of the complex motavizumab (colored green and magenta) with epitope cyan. (B) Back-view of the complex. (C) Side chains on the interface of the complex are shown in sticks.

Here, we describe a general and flexible computational approach to incorporate functional motifs into computationally designed proteins. Our strategy is readily applicable to wide variety of design problems without major limitations on the use of template topologies or different binding sites as long as there is a minimal degree (low resolution) of compatibility between them. We sought to fold a three-helical bundle of 120 residues around of a neutralizing epitope of RSV. The computational designs were characterized by a variety of biophysical methods and were shown to be folded and some with extremely high thermal stability. More importantly, the designs showed much higher affinities to the motavizumab Fab than the peptide[140], in the same range as the viral F protein[141]. The computational strategy together with the experimental validation implies that this bottom-up approach of folding a protein to ensure structural mimicry and support of a particular functional site might indeed be a useful and efficient.

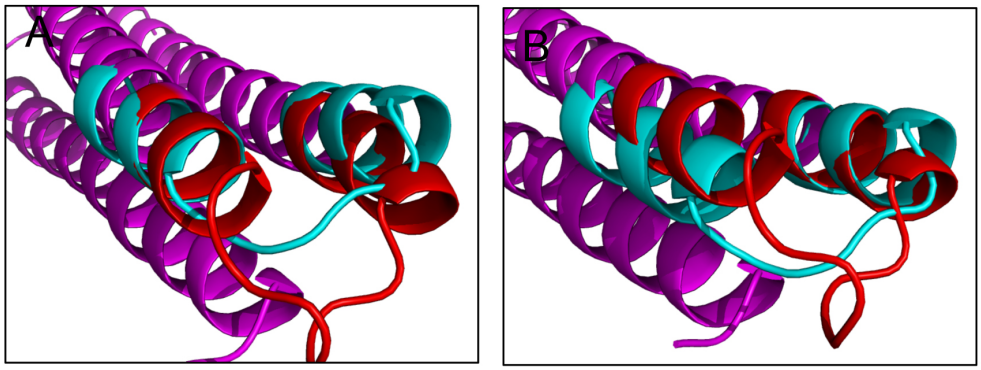


Figure 5.3: Structural differences between the epitope and the topology used as template to be folded. Target topology is shown in magenta, in red is the location where the epitope will be placed and in cyan is shown the epitope. The two structures are not superimposable due to different helix-helix spacing and orientation.

5.3 Material and Methods

5.3.1 Computational Procedure

We developed a general computational procedure as an extension to the Rosetta molecular modeling package[111, 77]. First, a binding site (PDBid: 3IXT, chain P)[140] and a target tridimensional topology (PDBid: 3LHP, chain S) are selected. The residue range of the target topology to be replaced by the binding site is defined. Second, the full atom coordinates of the binding site and the target topology are input and C-alpha distance constraints extracted from the target topology. Distances between the C-alpha atoms of residues with a sequence separation larger than 6 are collected outside the range where the binding site will be inserted. Third, an extended polypeptide chain is grown from the termini of the binding site atomic coordinates. Resulting in a centroid structure with the original dihedral angles of the target binding site and all remaining residues are in an extended conformation. Fourth, the extend chain is folded while the binding site dihedral angles are kept fixed acting as a folding nucleus. The folding stage is carried using structure prediction type techniques[147, 22, 148], where C-alpha distances collected from the target topology are used to restrict the conformational sampling. C-alpha distances of the target topology were

smoothed by allowing a 1.5 or 3 Å of standard deviation. Only structures with a backbone RMSD less than 5 Å relative to the target topology were selected for the full atom stage. Fifth, original side chain conformations of the binding site are recovered; typically 3 cycles of iterative design and full-atom optimization are performed. Sixth, an energy filter was employed and the best 50 designs were selected out of 10,000 for human inspection. Seventh, designs were further filtered according to several structural features Ramachandran score[111], Packing score[112] and buried unsatisfied atom counts. Additionally, a set of designs was also selected according to the average helical bend angle of the three helices as implemented in Helanal[149]. Eighth, surface residues of the initial topology were designed in each of the selected scaffolds, some mutations in the proteins cores were performed based on human inspection and to mutate overrepresented ALA to VAL/LEU or MET, and also to get rid of sporadic polar amino acids in the core, followed by a step of full-atom optimization[111]. The final filtering step was performed based on energy, packing, absence of steric clashes with the antibody and visual inspection.

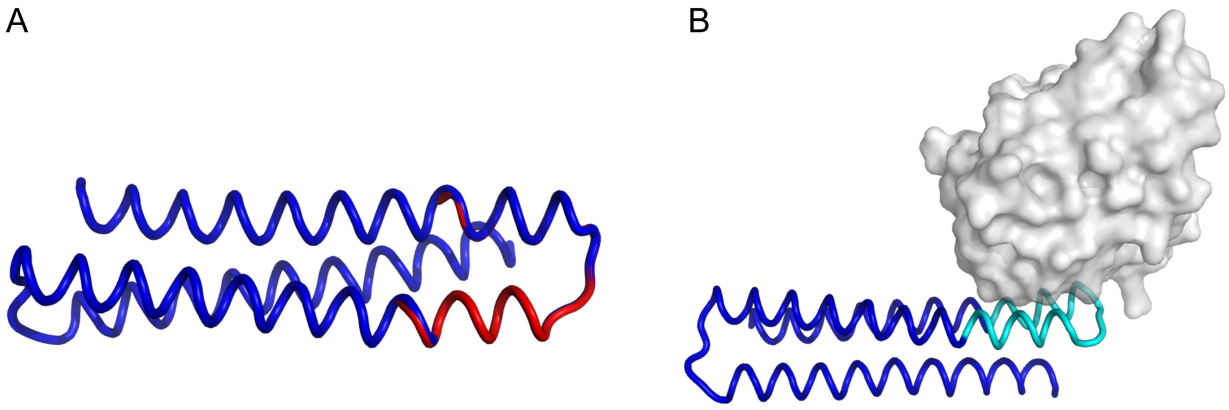


Figure 5.4: Structural features of a final FFL design. (A) Structural superposition of the epitope (colored in red) and the final FFL design (colored in blue). (B) FFL design (colored in blue with the epitope in cyan) in complex with Motavizumab (show in grey surface). The computational models of the FFL designs exhibit the adequate compatibility with the antibody, with no steric clashes or considerable extra contacts between scaffold and antibody.

5.3.2 Protein Expression and Purification

DNA segments encoding scaffold constructs were synthesized with optimized codon usage and RNA structure (Codon Devices, Genscript Corp.), subcloned into pET29 (EMD Biosciences) and transformed into Arctic Express^{TM}*E. coli* (Invitrogen). Single colonies from the transformation were grown overnight at 37 ^{o}C in starter cultures with Luria Broth (LB) plus Kanamycin (100 mg/ml). 10 mL of starter cultures were expanded into 1 L of LB plus Kanamycin (100 mg/mL) and incubated at 37 ^{o}C, 250 μM of IPTG were added to the culture when log phase was reached and incubated overnight to induce protein expression at 12 ^{o}C. Cultures were then pelleted and resuspended in start buffer (160 mM Imidazole, 4 M Sodium Chloride, 160 mM Sodium Phosphate), a tablet of protease inhibitor (Roche) was added and the cell suspension was frozen at -20 ^{o}C. The cell suspension was thawed and 10ml of 10x Bugbuster (Novagen), 50 μL of Benzonase Nucleases (Novagen) and 1.7 μL of rLysozyme (Novagen) were added to lyse the cells, the suspension was then gently tumbled in an orbital shaker for 20 minutes. Lysed cells were pelleted and the supernatant was filtered through a 0.2 micron filter (Millipore). Supernatants were tumbled with 5 mL of nickel-NTA resin (Fisher) for 1 hour at 4 ^{o}C. The resin was washed then rinsed with Wash Buffer (50 mM imidazole, 500 mM Sodium Chloride and 160 mM Sodium Phosphate) and eluted with elution buffer (250 mM Imidazole, 500 mM Sodium Chloride and 20 mM Sodium Phosphate). Fractions containing the construct of interest were combined and further purified by preparative size exclusion chromatography (SEC) on Superdex 75 16/60 (GE Healthcare) at room temperature in PBS or HBS. Fractions containing the protein of interest were collected, confirmed by SDS page and concentrated by ultrafiltration (Vivaspin, Bioexpress). Protein concentration was determined by measuring UV absorption signal at 280 nm and calculated from known amino acid composition extinction coefficient. To facilitate protein quantification in some of the designs was added the GLY-SER-TRP sequence on the C-terminus.

5.3.3 Biophysical Characterization

Solution oligomerization state and molecular weight

The monodispersivity and molecular weight were further assess by analytical SEC (Agilent, 1200 series) coupled to an on-line static light scatter (SLS)(miniDAWN TREOS, Wyatt). The molecular weight determination was performed with the ASTRA software (Wyatt).

Thermal stability

Solution thermostabilities (T_m) were determined by circular dichroism with an Aviv 62A DS spectrometer. Far-UV wavelength scans (190-260 nm) at concentrations ranging from 20 μM to 90 μM were collected in a 1mm path length cuvette. Temperature-induced protein denaturation was followed by change in ellipticity at 210 nm.

Nuclear magnetic resonance

One dimensional 1H were recorded at 25 ^{o}C on a Bruker DMX 500 MHz spectrometer. The designed proteins at concentrations ranging from 0.2 to 0.6 mM in Phosphate buffer (pH 7.4).

Surface plasmon resonance

All experiments were carried out on a BIACORE 2000 at 25 ^{o}C. Running buffer HBSEP (0.01 M HEPES pH 7.4, 0.15 M NaCl, 3 mM EDTA and 0.005% (v/v) Surfactant P20) (GE Healthcare). For analysis designs were immobilized on Research grade CM5 sensor chip. Designs were immobilized in three flow cells at similar levels (20 to 50 RUs) using Amine Coupling kit (GE Healthcare) according to the manufacturers manual, the fourth flow cell was used as reference cell and no design was immobilized. The analyte used was the Motavizumab Fab, all data were collected using two replicate injections for each concentration. The regeneration was performed with 10 seconds injection of H_3PO_4 (20x dilution) from Human Antibody Capture kit. Flow rates during the experiment were maintained at 50-100 μl/min. Data preparation and analysis were performed using Scrubber software

(version 2.0, BioLogic Software,Campbell, Australia). For kinetic analysis biosensor data were globally fit to a mass transport limited simple bimolecular binding model:

$$A_0 \underset{k_m}{\overset{k_m}{\rightleftharpoons}} A + B \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} AB$$

A_0 represents injected analyte

5.4 Results

5.4.1 Computational Design

The general computational method described is able to generate conformations and sequences where a binding or functional site works as a folding nucleus and the remaining polypeptide chain is folded around to support and stabilize perfectly the precise structure of the folding nucleus. We supplied an epitope of the RSV virus, a structural motif with biomedical relevance, to be the folding nucleus, and as template topology a three helix bundle redesigned in a previous work from our laboratory (Correia et al. Structure, Correia et al. submitted) that has been shown to be a stable protein. The chosen folding nucleus was crystallized in complex with Motavizumab Fab[140]. When bound to Motavizumab, the peptide adopts a helical hairpin conformation, which is suitable to become one of the helical hairpins of the 3-helix bundle. The replaced helical hairpin of the template topology had 4 residues more in the connecting loop and the RMSD of the backbone of the template and the folding nucleus was 2.6 Å.

In the initial folding stage low-resolution models (side chains represented as centroids) are generated, where the backbone of the folding nucleus is kept rigid and the remaining chain is folded employing C-alpha distance constraints. The side chain conformations from the residues in the folding nucleus are recovered before the design stage and are held fixed throughout the procedure maintaining all the atomic detail. The design stage is implemented as an iterative sequence-structure optimization protocol, where three steps of design[15] and full-atom minimization were performed[111]. The designs were filtered ac-

cording to Rosetta full atom energy, with the best 50 models by energy were further filtered according to general structural properties such as: Ramachandran score according the Ramachandran backbone dihedral angles distributions[111], counts of polar atoms buried in the surface of the protein and not involved in hydrogen bonds and packing assessment[112], all of this scores implemented as Rosetta. Additionally, the designed models were also filtered according a topological criterion selecting for the lowest average bend angle of the 3 helices[149]. The selected models were selected to test the influence of the overall straightness of the designed bundles. The surface positions were designed to be identical to the surface in the protein used as template topology. Finally, a step of manual design and all-atom optimization was employed to correct core-packing defects and remove any remaining polar amino acids from the core. The manual design was employed upon the observation that designs from the automated stage had alanines overrepresented and valines underrepresented, comparing to the template topology.

Given the nature of the iterative-design procedure the designed backbones were highly intolerable to mutations in the core, consequently in the manual design stage steric clashes were introduced by the performed mutations and a relax step was necessary to improve the full-atom energies. The relax step served also as a filtering step to the performed manual mutations, as a given mutation or set of mutations would only be accepted if the full atom energy recovered to the initial levels without creating major distortions in the helices local structure. Computational simulations were performed with a variety of conditions, and as mentioned, the filtering stage was also employed using different criterion (Table 5.1).

Among the many successes in computational design, several reports have describe the design of helical bundles but to our knowledge, the strategy of explicitly impose a folding nucleus in a computational calculation has not yet been explored. To validate the computational approach, eight designs were selected for experimental characterization. The diversity of the computational designs was observed both in terms of sequence and structure. Sequence-wise the differences are considerable, with the closest pair of designs differing by 8 mutations and the most distant pair differing by 42 mutations. The overall core is composed by 45 mutations and the number of mutations relative to the original protein used as the topological template ranged from 51 to 59 (Table 5.2, Figure 5.5). Structurally the

Table 5.1: Procedural details of each FFL design. The different parameters used to run the FFL simulations are summarized as well as the filtering criteria. The manual intervention stage is also summarized in terms of how many mutations were performed, the initial energy of the designs and the energy after a full-atom optimization step.

Molecule	Computational Algorithm		Filtering		Manual Intervention		
	SD(Å)[a]	BS design[b]	Energy + Composite Filters	Helix Bend[c]	Energy	Mutations	Energy post relax[d]
FFL_001	1.5	×	✓	×	-	-	-
FFL_002	3.0	✓	✓	×	-289	10	-276
FFL_003	3.0	✓	✓	×	-286	4	-293
FFL_004	3.0	✓	✓	×	-285	7	-291
FFL_005	1.5	✓	×	✓	-292	11	-287
FFL_006	3.0	✓	✓	×	-291	3	-290
FFL_007	1.5	✓	×	✓	-293	11	-285
FFL_008	1.5	✓	×	✓	-293	8	-286

[a] SD - standard deviation allowed to the constraints derived from target topology

[b] BS - Binding site design of the positions that are not in direct contact with the antibody

[c] Filtering criteria based on the helix bend angle as implemented in Helanal[149]

[d] Rosetta energy after the mutations have been performed and a step of full atom optimization

differences in backbone RMSD of the two most similar models is 0.5 Å and the furthest models are 3.1 Å away (Table 5.3). The RMSD differences to the template topology range from 1.84 to 2.91 Å. To confirm that the designed sequences indeed show an energetic preference for the conformation generated we performed a series of structural prediction[111] simulations where the designed sequences were folded from an extended chain procedure and no constraints were used. Six out of eight sequences showed a strong preference for conformations close ($\sim$ 2.5 Å away) to the designed conformation (Figure 5.10). In Rosetta structure prediction simulations some of the sequences also show an energetically favorable alternative conformation (Figure 5.10).

Table 5.2: Sequence diversity between designs and template topology. The numbers shown are the sequence mutations between each construct.

Molecule	FFL_001	FFL_002	FFL_003	FFL_004	FFL_005	FFL_006	FFL_007	FFL_008
T93	54	57	57	56	51	56	56	**59**
FFL_001	-	40	37	37	36	37	38	42
FFL_002		-	41	39	38	37	41	36
FFL_003			-	**8**	33	41	31	37
FFL_004				-	33	42	31	36
FFL_005					-	38	33	35
FFL_006						-	36	41
FFL_007							-	38

Table 5.3: Structural diversity between designs and template topology. RMSD calculations are show for the backbone atoms.

Molecule	FFL_001	FFL_002	FFL_003	FFL_004	FFL_005	FFL_006	FFL_007	FFL_008
T93	1.84	1.87	1.92	1.93	1.83	1.99	1.86	**2.91**
FFL_001	-	1.77	1.47	1.52	1.21	1.81	1.66	3.06
FFL_002		-	1.64	1.74	1.43	1.48	1.48	2.40
FFL_003			-	**0.53**	1.21	1.30	0.72	2.48
FFL_004				-	1.28	1.36	0.78	2.40
FFL_005					-	1.44	1.40	2.89
FFL_006						-	1.15	2.36
FFL_007							-	2.15

5.4.2 Folding and Stability

The designs selected for experimental characterization were generated by different simulations and filtering criteria. The variables were: residues within the binding site that were allowed to mutate; filtering based on the bend angle of the helical bundle; manual intervention. By having selected designs according different criteria we sought to learn what are the most robust filtering features and by including a design with no manual intervention we also wanted to compare a pure computational design with the ones where there was manual intervention.

From the eight designs experimentally tested, one (FFL_008) was not expressed in *E. coli*. The remaining seven were purified by immobilized metal affinity chromatography and

```
GSMSDRRKDLEERLDKLLEAAKNKEDKFKAAMRKRGQREERMKDWAKIARDEFEQFRKAV 60 FFL_003
GSMSDARKDLEERLDKLLEAAKNKMDKFKAAMRKRGQREERKKDWAKIVRDEFEQFRKAV 60 FFL_004
GSLSDIRKDAERRFDKLVEAVKNKLDKMKAALRKEGQQEERMKDLMKFMRKEVEQLRKAM 60 FFL_007
GSRSDMRKDAERRFDKFVEAAKNKFDKFKAALRKGDIKEERRKDMKKLARKEAEQARRAV 60 FFL_001
GSLSDVRKDVEKRIDKALEAFKNKMDKEKAAFRKDPPSEERRKDKKKEFREEREQVRKAI 60 FFL_002
GSFSDIRKDAEDRADKAFEAAKNKFDKIKAAIRKDWPSEERAKDLMKKARYEMEQARRAI 60 FFL_006
GSLSDLMKDLEKRFDKFMEAIKNKWDKVKAAFRKQEKGEERAKDMFKIFREELEQLRKAI 60 FFL_008
GSMSDIRKDLEERFDKLVEALKNKVDKMKAAFRKDQFHEERMKDWFKDLRKEVEQMRRAV 60 FFL_005
GSISDIRKDAEVRMDKAVEAFKNKLDKFKAAVRKVFPTEERIKDWLKIVRGEAEQARVAV 60 T93
** **  ** * * ** .** *** ** ***.**    *** **  *  * * ** * *:

RNFLSEALSKIN----DYPITNDDKKLTSNDAKKFDAEVAKKLEAFKADAEEAATQ--- 112 FFL_003
RNFLSEALSKIN----DYPITNDDKKLTSNDTKKFAAEVEKKLEAFKADVEEAATQ--- 112 FFL_004
RNFLSEALSKIN----DMPITNDDKKLISNDLKKYDAIAEKKLEAMKADVERMATQGSW 115 FFL_007
RNRLSELLSKIN----DMPITNDQKKLMSNDVLKFAAEAEKKIEALAADAEDKFTQGSW 115 FFL_001
RNVLSEALSKIN----DLPITNDKKKLVSNDVIKKVAEMKKKVELEVADVEKKVTQGSW 115 FFL_002
RNIESEALSKIN----DLPITNDQKKLASNDIIKEMARLFKKLEALMADIEILVTQ--- 112 FFL_006
RNALSEALSKIN----DLPITNDDKKLASNKAKKRAARVMKKVEAFIADVEAWKTQ--- 112 FFL_008
RNYASEALSKIN----DLPITNDDKKLASNDVLKLVAEVWKKLEAILADVEAWFTQ--- 112 FFL_005
RNVGRDANDKAAALGKDKEINWFDISQSLWDVQKLTDAAIKKIEAALADMEAWLTQ--- 116 T93
**   :  .*      *  *.  . .    .  *      **:*   ** *   **
```

Figure 5.5: Sequence alignment of the designs and sequence of the protein used as target topology (T93). Many of the surface residues of the target topology used were imposed in the designs.

size-exclusion chromatography (SEC). To investigate the multimer solution state the designs were concentrated up to 0.5 mM and analyzed by analytical SEC on-line with a static light scatter for molecular weight determination. Only one protein showed multiple molecular species (FFL_003) (data not shown) while the remaining six designs were monodisperse and with calculated molecular weights of 14 to 18 kDa, when the expected molecular weight of the monomeric designed proteins was approximately 15 kDa (Figure 5.6).

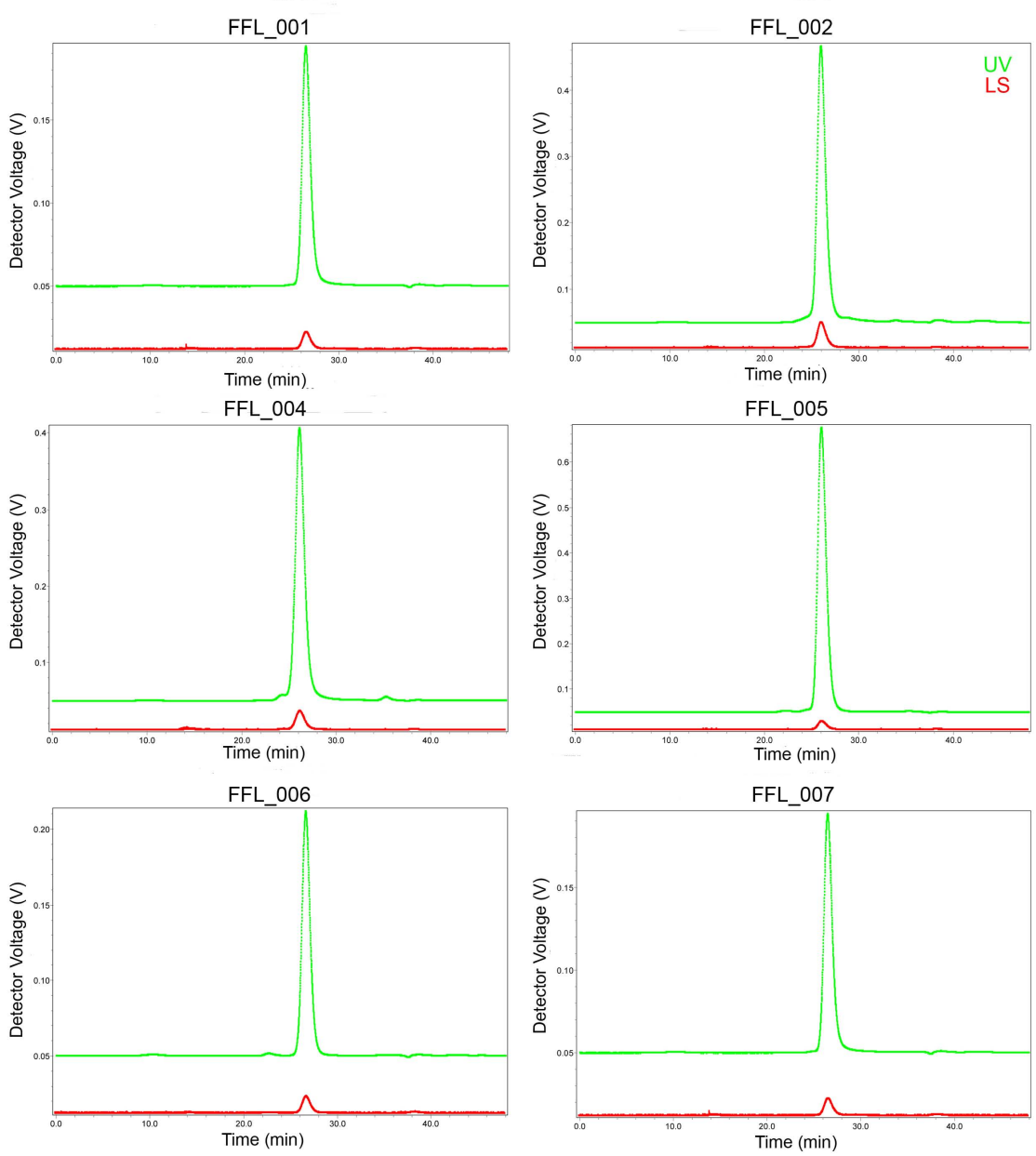


Figure 5.6: Characterization of the oligomeric state of the designs by SEC and SLS. All the molecules that showed a single monodisperse species had a molecular weight similar to the expected for a monomeric species, approximately 15 kDa. In green is the UV signal and in red is shown the light scatter signal.

Secondary structure content and thermal stability (T_m) of the monodisperse designs were investigated using CD spectroscopy. The CD spectra of the designs are typical for alpha-helical proteins showing two minima at 208 and 222 nm. Thermal stabilities ranged from 48 ^{o}C (FFL_002) to higher than 100 ^{o}C (FFL_005), the initial scaffold had a T_m of 79 ^{o}C. FFL_002 core sequence is a pure Rosetta design with no manual intervention, and FFL_005 shows a remarkable thermal stability with the wavelength spectra at 25 and 98 ^{o}C being very similar. Notably, FFL_003 and FFL_004 differ by only eight mutations, FFL_003 is not monodisperse and FFL_004 has a thermal stability higher than 85 ^{o}C (Figure 5.7 and Table 5.4).

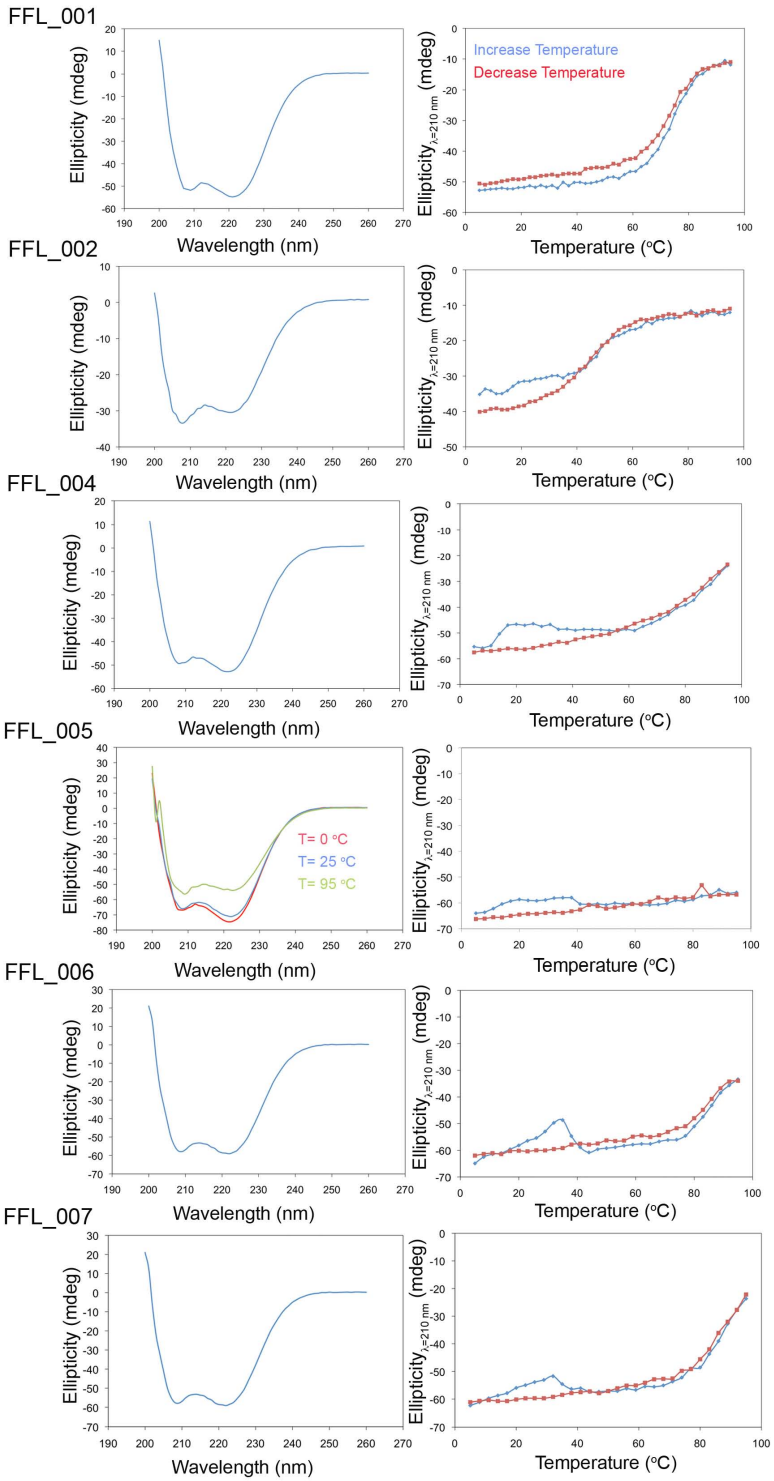


Figure 5.7: Secondary structure and thermal stability assessed by CD. On the left panel are shown wavelength scans, where the designs show typical spectra of all-helical proteins. On the right panel, thermal denaturation curves are shown and the designs unfold cooperatively and are reversible. FFL_005 is the exception, a very stable design that did not unfold by thermal denaturation, as shown by the wavelength scan at 95 ^{o}C.

To further assess the conformational dynamics of the designs proton NMR spectra were recorded. One-dimensional proton experiments show that the proteins are folded. Furthermore, is possible to observe some spectroscopic signatures of well behaved molecules, such as, the sharp narrow peaks observed on the amid proton region (9 to 6 ppm) observed for FFL_001 and well defined signal dispersion on the aliphatic region (1.0 and -1.0 ppm) in FFL_005. Contrarily, it is also possible to observe that the FFL_007 spectra does not exhibit any of the "good" spectroscopic signatures mentioned above[150] (Figure 5.8).

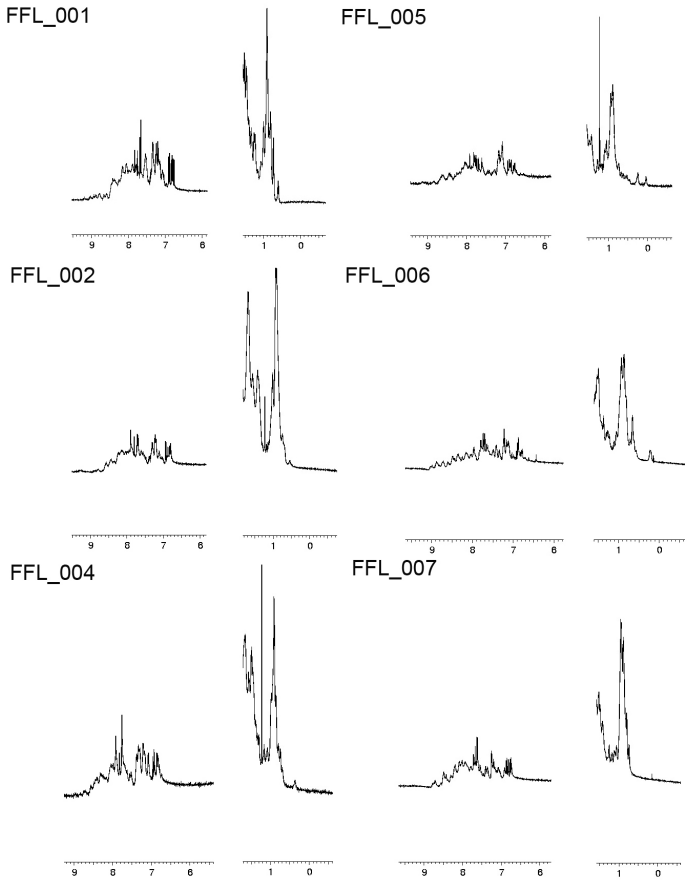


Figure 5.8: 1D ^{1}H-NMR spectroscopy of the FFL designs. In general the designs show good peak dispersion, signature properly folded molecules.

5.4.3 Binding to Motavizumab

Ultimately the aim of the computational method developed was to design molecules that provided the maximal structural support to the folding nucleus, stabilizing the exact conformation of the functional site that acted as a folding nucleus. We employed the developed computational methodology to a broadly neutralizing epitope of RSV that served as folding nucleus. To assess the validity of the designs we tested the affinities to the motavizumab using SPR. The designed scaffolds were immobilized on the chip and flowing motavizumab Fab as the analyte. The six designs showed extremely high affinites to the Fab, ranging from 795 pM to 29 pM. The tightest binder with the improvement in affinity coming from a slowest off rate ($2.3\text{x}10^{-3}\text{s}^{-1}$ for the highest affinity design $1.196\text{ x }10^{-4}\text{s}^{-1}$ for the lowest affinity design)(Figure 5.9 and Table 5.4). The reported affinity for the peptide epitope, used as a folding nucleus is 230 nM[140], the designs showed a $\sim$7900 fold improvement. We speculate that the increase in binding affinity observed on the designed scaffold comes from exact stabilization of the epitope. Binding of Motavizumab to peptide in solution requires a significant decrease in entropy of the flexible peptide, which is energetically unfavorable. Stabilization of the peptide by scaffolding eliminates that entropic barrier. While it is possible that the increase in affinity is due to additional contacts made by the scaffold outside the epitope region, the computational modeling suggests that is unlikely. Furthermore, the affinity of Motavizumab for the F protein of the RSV virus was reported previously as 35 pM, which is similar to several of the designed proteins. We speculate that this affinity is achieved by accurate mimicry of the binding interaction that occurs with the viral F protein and the antibody.

Table 5.4: Thermal stabilities and binding affinities of the FFL designs. Thermal stabilities were assessed by CD. K_Ds were determined between motavizumab Fab and the FFL designs.

Molecule	T_m (°C)	SPR		
		k_{on} ($M^{-1}s^{-1}$)	k_{off} (s^{-1})	k_{off}/k_{on} (pM)
FFL_001	75	3.99×10^{6}	1.19×10^{-4}	29.98
FFL_002	48	1.56×10^{6}	7.34×10^{-4}	469.9
FFL_004	>85	1.05×10^{6}	8.32×10^{-4}	795
FFL_005	>100	2.97×10^{6}	2.09×10^{-4}	70.3
FFL_006	>85	3.57×10^{6}	2.32×10^{-4}	651.9
FFL_007	>85	1.45×10^{6}	1.36×10^{-4}	94.1

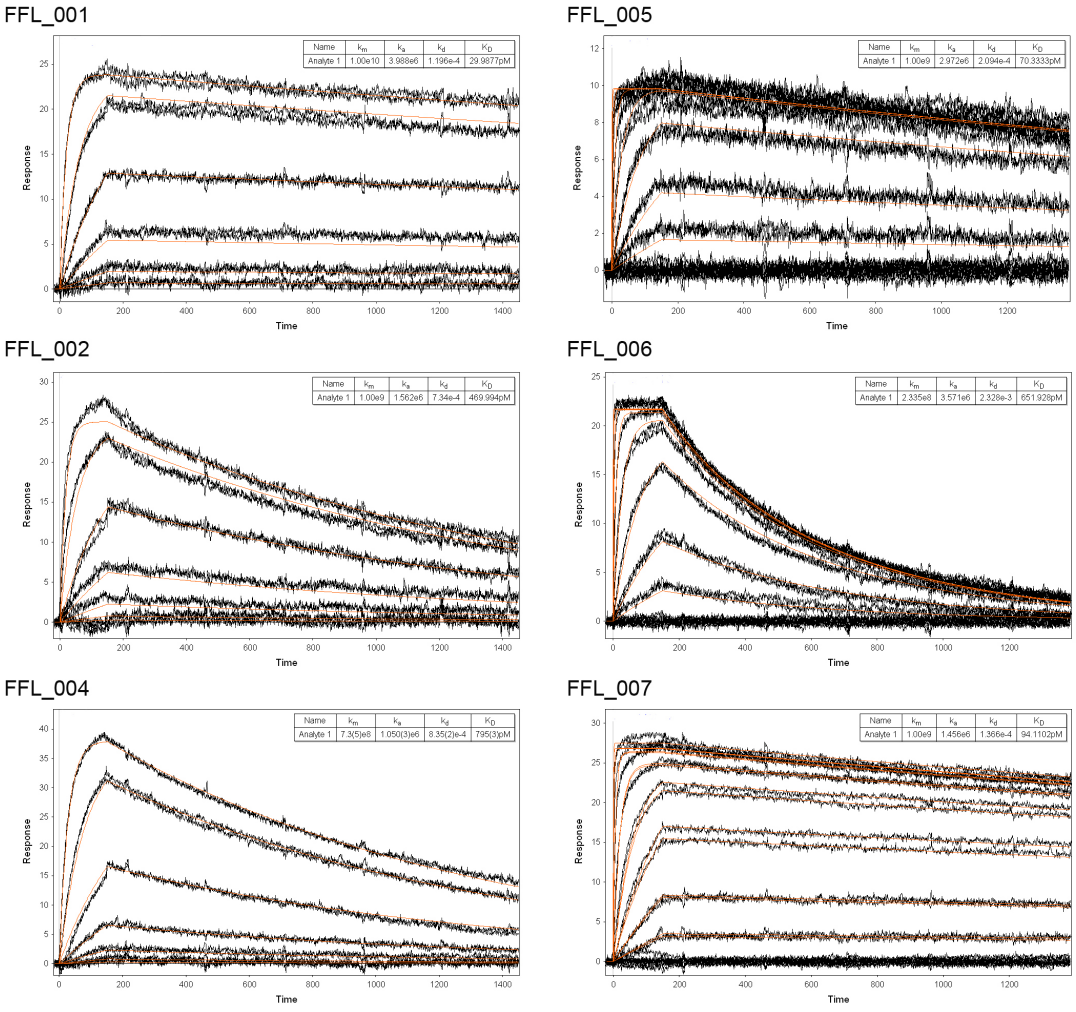


Figure 5.9: Binding affinities of the FFL designs to Motavizumab assessed by SPR. The epitope scaffolds were coupled to the biacore chip and Motavizumab Fab was used as analyte. The kinetic fits are shown in red.

5.5 Discussion

The data presented shows that we were able to rationally design new proteins with imposed functional constraints. Several of the designed proteins, bind with very high affinity, indicating a successful optimization of the functional constraint imposed. Simultaneously, the designs were highly soluble and thermally stable, showing that the folding free energy was also efficiently optimized.

The designed molecules are potentially relevant to vaccine research efforts for RSV a prevalent cause of illness among infants and children. The functional site incorporated is a well characterized epitope to which multiple neutralizing antibodies bind. The experimental characterization showed that the designed proteins achieved picomolar affinities to Motavizumab (highest 30 pM; lowest 750 pM). The affinities of the designs are in the same order of the viral F protein; previous epitope scaffolding effort (Mclellan and Correia in preparation), based in side-chain grafting (Correia et al., Ofek et al.), was never able to attain affinities observed in the latest generation. In designed immunogens the affinity for the antibody to be re-elicited one of the most important features. Remarkably, the designs showed a high amount of sequence variability for the protein core, in some cases with high thermal stability.

The methodology described is readily applicable to a myriad of protein design problems, such as binding site transplantation or enzyme design. For proof of concept, we utilized a previously observed fold in nature to extract the C-alpha constraints; nevertheless, this approach can be extended to folds not yet observed in nature. The binding site used was a contiguous segment but the procedure is also able to deal with discontinuous functional sites composed by two or more segments, such as catalytic triads. At this point the major limitation we foresee is the size of the protein to be folded, approximately 150 residues should be about the upper limit. This limitation is due to computational algorithmic issues related to the larger space and rugged conformational landscapes available to larger polypeptide chains.

The design of the FFL proteins is a demonstration that the Rosetta energy function was able to optimize folding free energy in the presence of functional constraints. Our work

shows that the field of protein design is maturing at a considerable pace. Here we proposed a strategy to leverage the accuracy of energy functions together with realistic sampling of the conformational space to computationally design functional proteins.

5.6 Acknowledgements

Bruno Emanuel Correia (B.E.C.) wrote the Fold From Loops code and designed the FFL proteins. B.E.C. and Oleksander Kalyuzhniy performed biophysical characterization of the designs. Alex Schroeter performed protein expression and purification. Joseph Jardine commented on the manuscript and is performing crystallization experiments. William Ray Schief and David Baker supervised research.

5.7 Supplemental Material

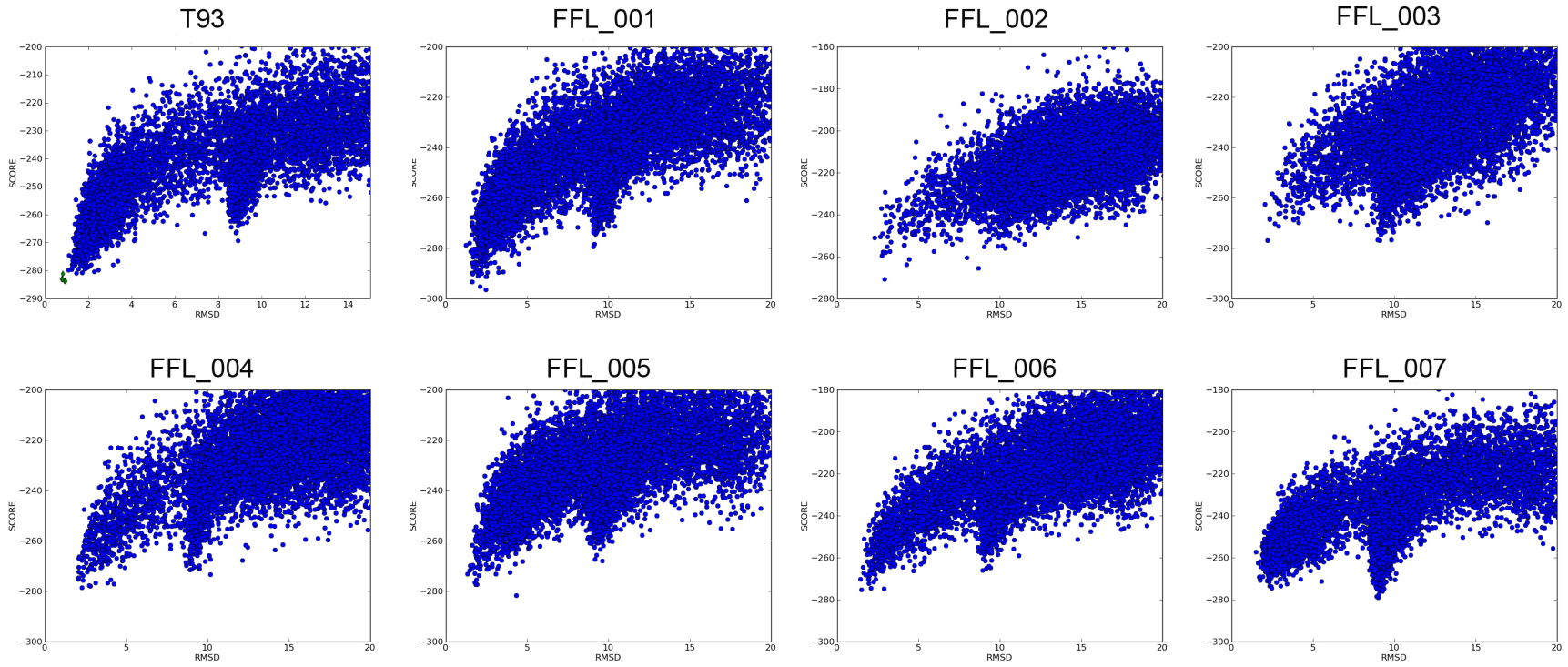


Figure 5.10: Forward structure prediction with the sequences extracted from the FFL procedure. The green diamonds in the T93 plot represent the native structure after Rosetta full atom optimization.

Chapter 6

FINAL DISCUSSION

Throughout my graduate studies the main focus was to develop and apply computational design methodologies in order to create molecules with biological and therapeutic applications. In this work I described the concept of designing "epitope-scaffolds" applied to two different viruses (HIV and RSV), using both fixed and flexible backbone approaches. Several contributions were made for the advance of the fields of computational and immunogen design:

1. Side chain grafting was used to design a panel of 4E10 epitope-scaffolds;

2. Immunization studies showed that the 4E10 epitope-scaffolds elicited epitope specific responses;

3. Flexible backbone design methodologies were developed and applied to the removal of an immunodominant domain;

4. A discontinuous epitope was grafted onto a non-related protein and computational design was used to leverage *in vitro* evolution;

5. Development of a novel method with high backbone freedom - Fold From Loops - to fold and design a protein scaffold around a binding or functional site.

6.1 Design of 4E10 epitope-scaffolds using side chain grafting

As reviewed in the General Introduction, several methods existed to place multiple side chains in particular geometrical arrangements. Our newly developed side chain grafting method searches for similar backbones to the epitope of interest, checks structural compatibility of the potential scaffolds with antibody binding and then critical side chains for the

interaction are grafted. The results obtained show that we were able to transplant contiguous epitopes, achieving exquisite atomic mimicry and improved antibody affinities in comparison with the epitope peptide. Our rational approach enabled the design of candidate immunogens with a considerable chance of success.

Side chain grafting was one of the first structure-based strategies employed to transplant linear motifs of broadly neutralizing epitopes from HIV. The epitope-scaffolds represent advances in the field of immunogen design, where typically many of the candidates for immunization are the original proteins subunits from pathogens. Often vaccine candidate molecules are generated in a semi-empirical fashion where structural and biochemical properties are often compromised. The epitope-scaffolds also surpass the molecular tricks that evolved on the original proteins from pathogens to escape the immune system. Immunization studies were carried with the first generation of side chain grafting epitope-scaffolds. Sera from immunized rabbits cross-reacted with the 4E10 epitope. Furthermore, experiments using alanine mutants of the 4E10 epitope showed that 4E10-like antibodies were re-elicited.

Our work is one of the few in the literature reporting elicitation of 4E10-like antibodies, as this epitope is notoriously poorly immunogenic[57, 65, 66]. The pitfall was the lack of neutralization capability of the sera elicited by the epitope-scaffolds. The central question of the project became: if the re-elicited sera specificity matches the 4E10 MAb specificity, why is it not able to neutralize the virus? Two main future directions are plausible to improve the lack of neutralization:

a) Importance of the CDRH3 of the 4E10 antibody in virus neutralization. Several reports had dissected the importance of the CDRH3 for neutralization[61, 62, 63, 64] and demonstrated that a few mutations on the tip of the CDRH3 (Figure 6.1) reduce the 4E10 neutralization potency without impacting peptide binding. There has been some evidence that these mutants reduce the interaction of 4E10 antibody with lipid membranes, suggesting that this interaction is important for viral neutralization. The designed epitope-scaffolds only recreate the 4E10 epitope portion, which seems to be required but insufficient to achieve neutralization through the 4E10 antibody. A future route is to create epitope-scaffolds where explicit atomic interactions are required with the CDRH3 of the antibody.

This extra requirement can be implemented in the matching stage, to search for scaffolds with pre-existing patches able to interact productively with the 4E10 CDRH3. More aggressive strategies using flexible backbone design approaches to build *de novo* interaction patches can also be utilized. In principle, the success of the CDRH3-complementarity designs can be assessed by comparing the affinity of such epitope-scaffolds to wild-type 4E10 versus CDRH3 mutant antibodies; a decrease in affinity to the mutant 4E10 would indicate that the CDRH3-complementarity designs were successful.

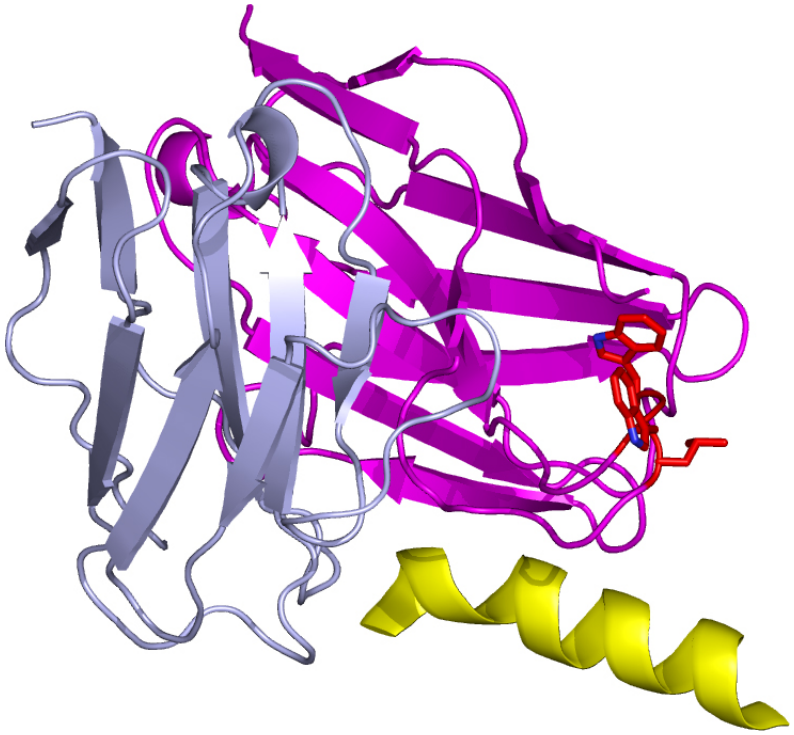


Figure 6.1: 4E10 CDRH3 residues impact virus neutralization. Residues on the tip of the CDRH3 loop are colored in red and the peptide epitope is colored in yellow.

b) Interaction with the naive germline antibodies. It is hypothesized that to stimulate potent and specific antibody responses to a particular antigen, one of the key steps is to interact with naive germline antibodies from the B-cell repertoire. As mentioned, the epitope-scaffolds were strictly designed to bind to the mature 4E10 antibody; and it was observed in a recent study that germline predecessors of HIV neutralizing antibodies where

unable to bind to Env proteins (gp120-gp41) from HIV [151]. Given the ability to trace back naive antibody germeline sequences[152], we can employ structure-based strategies to model naive antibodies and design epitope-scaffolds able to bind germline variants. Eventually, this effort could be extended aiming to engineer scaffolds capable of binding both variants of the 4E10 antibody, becoming an interesting challenge of designing dual-specificity. Nevertheless, it seems that to accomplish such design the use of *in vitro* evolution and selection techniques would be desirable.

Despite the lack of neutralization, the epitope-scaffolds were a test to our knowledge of vaccine design and which features need to be encoded in efficient immunogens.

6.2 Flexible Backbone design to engineer structure and immunogenic properties of epitope-scaffolds

At the beginning of my project, few computational designs where backbone flexibility was employed had been reported, and mainly these designs aimed for structural recapitulation. Flexible backbone design was employed to remove an extraneous domain of an epitope-scaffold for 4E10 antibody binding. Beyond the structural objectives, the domain-trimmed scaffolds aimed to improve immunogenic properties of a 4E10 epitope-scaffold designed previously using side chain grafting. I devised a protocol that performs backbone sampling with pre-defined secondary structure, which was combined with RosettaDesign and finalized with a filtering stage to select designs with native-like features. The flexible backbone procedure yielded several designs that were soluble and with improved thermal stability relative to the parent molecule. One of the designs was crystallized and showed close agreement with the computational model. During the maturation of our method two other articles reported the use of flexible backbone design[48, 49]. Nevertheless, our strategy still remains general and adequate to several applications that might not be straightforward with the other reported approaches:

a) Hu et al.[48] reported the design of a new loop conformation. The starting backbones that were designed designed were selected from a structural database of loops with length 12. In our flexible backbone protocol, backbones were generated by fragment assembly conferring an extra layer of conformational sampling. Possibly, future design problems might

require the creation of backbone conformations not yet observed in the structural space available, hence assembling fragments to create structural diversity might be necessary.

b) Murphy and coworkers [49] also designed backbones generated from fragment assembly. But they used a pre-placed side chain central to the catalytic activity, constraining the loop conformations sampled. The loop built was 5 residues long, while in our case helix-loop segments with 16 and 17 residues were built, and 24 residues were designed in a protein with 120 residues.

Sera from immunizations with the parent scaffold cross-reacted strongly with the scaffold itself. Further analysis of the parent scaffold's sera showed weak cross-reactivity with a trimmed epitope-scaffold. This result confirmed that the majority of the antibodies in the sera were directed to the removed domain. The domain-trimmed scaffold lacks an immunodominant domain that might hinder epitope specific responses. Domain trimming and resurfacing (also described in Chapter 3) are valid structure-based strategies to generate panels of epitope-scaffolds that can be used for multiple applications. Frequently in immunization experiments, multiple antigens are used sequentially or combined to stimulate specific antibody responses, therefore variants of the epitope-scaffolds might be appropriate to be used in such immunization schemes. Additionally, these panels of optimized epitope-scaffolds are valuable reagents to dissect epitope specificities of polyclonal sera.

Methodologically, it would be important to test the protocol on the design of other secondary structure motifs, to further evaluate how general it can be. The principles developed in this domain trim problem served as a proxy to deal with more complex design problems.

6.3 Grafting of a complex epitope using computationally inspired directed libraries

Many of the epitopes and protein binding sites are not contiguous in sequence, but are comprised within multiple discontinuous segments of the protein, coming together in the tridimensional structure to form the interaction site. Given the necessity of scaffolding complex epitopes the Multigraft protocol was developed. Multigraft was applied to the b12 epitope from gp120, and we sought to transplant a "minimimal" b12 epitope were only the 2 core loops responsible by most of the contacts were considered. The structural complexity

of this design and the lack of success of the computational approach, suggested the need for high-throughput screening and selection of designs with affinity for the b12 antibody.

A yeast display platform was used to screen directed libraries that were assembled according to sequence profiles generated by computational design calculations. Here, I devised an iterative flexible backbone design protocol where individual segments of the scaffold were optimized and then recombined. The rational was to first stabilize the local conformation of each loop and after the recombination step, optimize the correct spatial orientation between them. A naive library allowing the 20 amino acids in all the positions hypothesized to be relevant for the stabilization of the 2 loops would span a sequence space of size approximately 10^{21}; a Rosetta library including all the amino acids observed in computational designs would have the size of 10^{17}; finally, after a strict filtering based on structural features of the computational models the library size was reduced to 10^{12}. In order to enable the screening using yeast display, the library was split in two, each targeting one of the loops of the epitope. We achieved an improvement of 30,000 fold over the initial computational design. The best clones selected bound in the same affinity range of the original gp120-b12 interaction. To our knowledge, this was the first successful transplantation of a complex binding site onto a non-related protein. The results suggest that computational sequence-structure ensembles may carry useful information to guide the construction of directed libraries, particularly useful if aggressive structural intervention is required.

Some groups started to explore the strategy of designing computationally inspired libraries [123, 124], but in some of these reports the aim is to recapitulate *in silico* results of selection experiments[153]. In our case the challenge was more demanding, since the starting library was designed from computational calculations. Utilizing high-throughput selection techniques has many advantages over the typical expression-purification-characterization of tens to hundreds of different constructs. It is possible to screen many more sequences in a combinatorial fashion increasing the chances of success, therefore providing valuable input to benchmark and improve the design methodologies. Another advantage of using computational guided libraries is that, in principle the chances of evolving clones with binding sites that were not pre-determined is greatly reduced, often unexpected results occur in random evolution experiments.

Structural characterization of successful constructs is still in progress and it will be essential to evaluate and improve the current computational protocol. In future implementations, we aim to make the protocol fully automated. It would be particularly interesting to implement an automated recombination step, where multiple Rosetta trajectories running independently in different nodes of a computational cluster could communicate and recombine their best solutions to further diversify the search space.

6.4 Transplantation and stabilization of functional sites by folding and design a protein around structural motifs

One of the major limitations of the scaffolding methods described above (Side chain grafting and Multigraft) is that both are critically dependent on a matching stage to find compatible scaffolds with the structural motif to be transplanted. During my PhD project I implemented and developed a computational method named Fold From Loops (FFL). Essentially, the procedure takes a structural motif and a protein topology, and by extracting distance constraints from the topology is able to fold a similar structure around the motif of interest. Two main fundamental differences between FFL and the methods previously describe:

a) FFL only requires a topological description of the protein to be folded disregarding atomic level details that both side chain and backbone grafting matching consider;

b) in FFL side chains or backbones are not being imposed to pre-existing protein structures, but instead most of the protein is flexible allowing proper accommodation and structural stabilization of the fixed functional site.

FFL was tested on a broadly neutralizing epitope from RSV which is structurally defined as a helical hairpin. A three-helix bundle was used as template for the topology to be folded. The best designs showed low pM affinities towards the cognate antibody (Motavizumab). In contrast to the nM afinities of previous scaffolds designed with an extension of the side chain grafting method (Mclellan and Correia, in preparation). So far, the results suggest that the FFL method is promising for the transplant of structural motifs.

The FFL protocol represents an advance in protein design algorithms, where functional and structural constraints are incorporated in the computational simulation. These constraints bias the overall structure to provide support to the functional site. It is also a long

waited development to the design exclusively for structural purposes, such as the design of the right handed coiled coil[19] and the new globular topology (Top7)[21]. For further development and validation of the FFL computational methodology, structural characterization is essential and is work in progress. For future applications, an appealing and important challenge is to fold proteins around two or three discontinuous protein segments, allowing the design of more diverse protein functions. The FFL algorithm could be a valid alternative to current enzyme design algorithms that place catalytic side chains onto pre-existing backbones without allowing any flexibility[30, 31, 32].

Overall in this thesis several advances are described, both in terms of immunogen design and methodological developments of general protein design. It is advantageous to have multiple computational design strategies available, providing appropriate alternatives to approach distinct design problems. In immunogen design, there is a broad future perspective for the developed methods, given that the active discovery of new broadly neutralizing antibodies provides novel templates for structure-based approaches. The methods described here will certainly benefit from parallel efforts of the scientific community, in particular on the development of more accurate energy functions and more realistic sampling methods. These improvements will contribute to the design of proteins with greater confidence, better success rates and the engineering of the most useful and creative protein functions.

BIBLIOGRAPHY

[1] K. Eric Drexler. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proceedings of the National Academy of Sciences of the United States of America*, 78(9):5275 –5278, 1981.

[2] Carl Pabo. Molecular technology: Designing proteins and peptides. *Nature*, 301(5897):200–200, 1983.

[3] B Gutte, M Dumigen, and E Wittschieber. Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids. *Nature*, 281(5733):650–655, October 1979. PMID: 551285.

[4] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucl. Acids Res.*, 28(1):235–242, January 2000.

[5] J W Ponder and F M Richards. Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, 193(4):775–791, February 1987. PMID: 2441069.

[6] H W Hellinga and F M Richards. Construction of new ligand binding sites in proteins of known structure. i. computer-aided modeling of sites with pre-defined geometry. *Journal of Molecular Biology*, 222(3):763–85, December 1991. PMID: 1749000.

[7] H W Hellinga, J P Caradonna, and F M Richards. Construction of new ligand binding sites in proteins of known structure. II. grafting of a buried transition metal binding site into escherichia coli thioredoxin. *Journal of Molecular Biology*, 222(3):787–803, December 1991. PMID: 1660933.

[8] B. Kuhlman and D. Baker. Exploring folding free energy landscapes using computational protein design. *Curr Opin Struct Biol*, 14(1):89–95, 2004.

[9] D. N Bolon, J. S Marcus, S. A Ross, and S. L Mayo. Prudent modeling of core polar residues in computational protein design. *J Mol Biol*, 329(3):611–22, 2003.

[10] Shannon A Marshall, Chantal S Morgan, and Stephen L Mayo. Electrostatics significantly affect the stability of designed homeodomain variants. *Journal of Molecular Biology*, 316(1):189–199, February 2002. PMID: 11829512.

[11] Martin Sagermann, Lars-Gran Mrtensson, Walter A Baase, and Brian W Matthews. A test of proposed rules for helix capping: implications for protein design. *Protein Science: A Publication of the Protein Society*, 11(3):516–521, March 2002. PMID: 11847274.

[12] B. I Dahiyat and S. L Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–7, 1997.

[13] B. I Dahiyat and S. L Mayo. Protein design automation. *Protein Sci*, 5(5):895–903, 1996.

[14] P Koehl and M Levitt. De novo protein design. i. in search of stability and specificity. *Journal of Molecular Biology*, 293(5):1161–1181, November 1999. PMID: 10547293.

[15] B Kuhlman and D Baker. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America*, 97(19):10383–8, September 2000. PMID: 10984534.

[16] K Raha, A M Wollacott, M J Italia, and J R Desjarlais. Prediction of amino acid sequence from structure. *Protein Science: A Publication of the Protein Society*, 9(6):1106–1119, June 2000. PMID: 10892804.

[17] Alfonso Jaramillo, Lorenz Wernisch, Stphanie Hry, and Shoshana J Wodak. Folding free energy function selects native-like protein sequences in the core but not on the surface. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21):13554–13559, October 2002. PMID: 12368470.

[18] A. Su and S. L Mayo. Coupling backbone exibility and amino acid sequence selection in protein design. *Protein Sci*, 6(8):1701–7, 1997.

[19] Pehr B. Harbury, Joseph J. Plecs, Bruce Tidor, Tom Alber, and Peter S. Kim. High-Resolution protein design with backbone freedom. *Science*, 282(5393):1462–1467, November 1998.

[20] F. H. C Crick. The packing of alpha-helices: simple coiled-coils. *Acta Crystallographica*, 6(8-9):689697, September 1953.

[21] B. Kuhlman, G. Dantas, G. C Ireton, G. Varani, B. L Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–8, 2003.

[22] K T Simons, C Kooperberg, E Huang, and D Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, April 1997. PMID: 9149153.

[23] John Karanicolas and Brian Kuhlman. Computational design of affinity and specificity at protein-protein interfaces. *Current Opinion in Structural Biology*, 19(4):458–463, August 2009.

[24] James J Havranek and Pehr B Harbury. Automated design of specificity in molecular recognition. *Nature Structural Biology*, 10(1):45–52, January 2003. PMID: 12459719.

[25] Gevorg Grigoryan and Amy E. Keating. Structure-based prediction of bZIP partnering specificity. *Journal of Molecular Biology*, 355(5):1125–1142, February 2006.

[26] Julia M Shifman and Stephen L Mayo. Modulating calmodulin binding specificity through computational protein design. *Journal of Molecular Biology*, 323(3):417–423, October 2002. PMID: 12381298.

[27] J. Reina, E. Lacroix, S. D Hobson, G. Fernandez-Ballester, V. Rybin, M. S Schwab, L. Serrano, and C. Gonzalez. Computer-aided design of a PDZ domain to recognize new target sequences. *Nat Struct Biol*, 9(8):621–7, 2002.

[28] Justin Ashworth, James J Havranek, Carlos M Duarte, Django Sussman, Raymond J Monnat, Barry L Stoddard, and David Baker. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, 441(7093):656–659, June 2006. PMID: 16738662.

[29] Alexandre Zanghellini, Lin Jiang, Andrew M Wollacott, Gong Cheng, Jens Meiler, Eric A Althoff, Daniela Rthlisberger, and David Baker. New algorithms and an in silico benchmark for computational enzyme design. *Protein Science: A Publication of the Protein Society*, 15(12):2785–2794, December 2006. PMID: 17132862.

[30] D. Rothlisberger, O. Khersonsky, A. M Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L Gallaher, E. A Altho, A. Zanghellini, O. Dym, S. Albeck, K. N Houk, D. S Tawk, and D. Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–5, 2008.

[31] Lin Jiang, Eric A Althoff, Fernando R Clemente, Lindsey Doyle, Daniela Rthlisberger, Alexandre Zanghellini, Jasmine L Gallaher, Jamie L Betker, Fujie Tanaka, Carlos F Barbas, Donald Hilvert, Kendall N Houk, Barry L Stoddard, and David Baker. De novo computational design of retro-aldol enzymes. *Science (New York, N.Y.)*, 319(5868):1387–1391, March 2008. PMID: 18323453.

[32] Justin B Siegel, Alexandre Zanghellini, Helena M Lovick, Gert Kiss, Abigail R Lambert, Jennifer L St Clair, Jasmine L Gallaher, Donald Hilvert, Michael H Gelb, Barry L Stoddard, Kendall N Houk, Forrest E Michael, and David Baker. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science (New York, N.Y.)*, 329(5989):309–313, July 2010. PMID: 20647463.

[33] K Hornischer and H Blcker. Grafting of discontinuous sites: a protein modeling strategy. *Protein Engineering*, 9(11):931–939, November 1996. PMID: 8961346.

[34] D P Hearst and F E Cohen. GRAFTER: a computational aid for the design of novel proteins. *Protein Engineering*, 7(12):1411–1421, December 1994. PMID: 7716151.

[35] Sen Liu, Shiyong Liu, Xiaolei Zhu, Huanhuan Liang, Aoneng Cao, Zhijie Chang, and Luhua Lai. Nonnatural proteinprotein interaction-pair design by key residues grafting. *Proceedings of the National Academy of Sciences*, 104(13):5330–5335, March 2007.

[36] H Domingues, D Cregut, W Sebald, H Oschkinat, and L Serrano. Rational design of a GCN4-derived mimetic of interleukin-4. *Nature Structural Biology*, 6(7):652–656, July 1999. PMID: 10404222.

[37] Samuel K Sia and Peter S Kim. Protein grafting of an HIV-1-inhibiting epitope. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17):9756–9761, August 2003. PMID: 12913122.

[38] Bin Qian, Srivatsan Raman, Rhiju Das, Philip Bradley, Airlie J McCoy, Randy J Read, and David Baker. High-resolution structure prediction and the crystallographic phase problem. *Nature*, 450(7167):259–64, November 2007. PMID: 17934447.

[39] Chu Wang, Philip Bradley, and David Baker. Protein-protein docking with backbone flexibility. *Journal of Molecular Biology*, 373(2):503–519, October 2007. PMID: 17825317.

[40] Ian W Davis and David Baker. RosettaLigand docking with full ligand and receptor flexibility. *Journal of Molecular Biology*, 385(2):381–392, January 2009. PMID: 19041878.

[41] Srivatsan Raman, Oliver F. Lange, Paolo Rossi, Michael Tyka, Xu Wang, James Aramini, Gaohua Liu, Theresa A. Ramelot, Alexander Eletsky, Thomas Szyperski, Michael A. Kennedy, James Prestegard, Gaetano T. Montelione, and David Baker. NMR structure determination for larger proteins using Backbone-Only data. *Science*, 327(5968):1014–1018, February 2010.

[42] Frank DiMaio, Michael D Tyka, Matthew L Baker, Wah Chiu, and David Baker. Refinement of protein structures into low-resolution density maps using rosetta. *Journal of Molecular Biology*, 392(1):181–190, September 2009. PMID: 19596339.

[43] Summer B Thyme, Jordan Jarjour, Ryo Takeuchi, James J Havranek, Justin Ashworth, Andrew M Scharenberg, Barry L Stoddard, and David Baker. Exploitation of binding energy for catalysis and design. *Nature*, 461(7268):1300–1304, October 2009. PMID: 19865174.

[44] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods*, 7(4):291–294, April 2010. PMID: 20190761.

[45] T Lazaridis and M Karplus. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *Journal of Molecular Biology*, 288(3):477–487, May 1999. PMID: 10329155.

[46] Tanja Kortemme, Alexandre V Morozov, and David Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology*, 326(4):1239–1259, February 2003. PMID: 12589766.

[47] R L Dunbrack and F E Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science: A Publication of the Protein Society*, 6(8):1661–1681, August 1997. PMID: 9260279.

[48] X. Hu, H. Wang, H. Ke, and B. Kuhlman. High-resolution design of a protein loop. *Proc Natl Acad Sci U S A*, 104(45):17668–73, 2007.

[49] Paul M Murphy, Jill M Bolduc, Jasmine L Gallaher, Barry L Stoddard, and David Baker. Alteration of enzyme specificity by computational loop remodeling and design. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9215–9220, June 2009. PMID: 19470646.

[50] D. R. Burton. Antibodies, viruses and vaccines. *Nat Rev Immunol*, 2(9):706–13, September 2002.

[51] G. Frey, H. Peng, S. Rits-Volloch, M. Morelli, Y. Cheng, and B. Chen. A fusion-intermediate state of HIV-1 gp41 targeted by broadly neutralizing antibodies. *Proc Natl Acad Sci U S A*, 105(10):3739–44, March 2008.

[52] L. Stamatatos, L. Morris, D. R. Burton, and J. R. Mascola. Neutralizing antibodies generated during natural HIV-1 infection: good news for an HIV-1 vaccine? *Nat Med*, 15(8):866–70, August 2009.

[53] A. K. Kashyap, J. Steel, A. F. Oner, M. A. Dillon, R. E. Swale, K. M. Wall, K. J. Perry, A. Faynboym, M. Ilhan, M. Horowitz, L. Horowitz, P. Palese, R. R. Bhatt, and R. A. Lerner. Combinatorial antibody libraries from survivors of the turkish H5N1 avian influenza outbreak reveal virus neutralization strategies. *Proc Natl Acad Sci U S A*, 105(16):5986–91, April 2008.

[54] Y. Okuno, Y. Isegawa, F. Sasao, and S. Ueda. A common neutralizing epitope conserved between the hemagglutinins of influenza a virus h1 and h2 strains. *J Virol*, 67(5):2552–8, May 1993.

[55] J. Sui, W. C. Hwang, S. Perez, G. Wei, D. Aird, L. M. Chen, E. Santelli, B. Stec, G. Cadwell, M. Ali, H. Wan, A. Murakami, A. Yammanuru, T. Han, N. J. Cox, L. A. Bankston, R. O. Donis, R. C. Liddington, and W. A. Marasco. Structural and functional bases for broad-spectrum neutralization of avian and human influenza a viruses. *Nat Struct Mol Biol*, 16(3):265–73, March 2009.

[56] M. Throsby, E. van den Brink, M. Jongeneelen, L. L. Poon, P. Alard, L. Cornelissen, A. Bakker, F. Cox, E. van Deventer, Y. Guan, J. Cinatl, J. ter Meulen, I. Lasters, R. Carsetti, M. Peiris, J. de Kruif, and J. Goudsmit. Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory b cells. *PLoS One*, 3(12):e3942, 2008.

[57] J. M. Binley, E. A. Lybarger, E. T. Crooks, M. S. Seaman, E. Gray, K. L. Davis, J. M. Decker, D. Wycuff, L. Harris, N. Hawkins, B. Wood, C. Nathe, D. Richman, G. D. Tomaras, F. Bibollet-Ruche, J. E. Robinson, L. Morris, G. M. Shaw, D. C. Montefiori, and J. R. Mascola. Profiling the specificity of neutralizing antibodies in a large panel of plasmas from patients chronically infected with human immunodeficiency virus type 1 subtypes b and c. *J Virol*, 82(23):11651–68, December 2008.

[58] G. Stiegler, R. Kunert, M. Purtscher, S. Wolbank, R. Voglauer, F. Steindl, and H. Katinger. A potent cross-clade neutralizing human monoclonal antibody against a novel epitope on gp41 of human immunodeficiency virus type 1. *AIDS Res Hum Retroviruses*, 17(18):1757–65, December 2001.

[59] M. B. Zwick, A. F. Labrijn, M. Wang, C. Spenlehauer, E. O. Saphire, J. M. Binley, J. P. Moore, G. Stiegler, H. Katinger, D. R. Burton, and P. W. Parren. Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *J Virol*, 75(22):10892–905, November 2001.

[60] Z. Y. Sun, K. J. Oh, M. Kim, J. Yu, V. Brusic, L. Song, Z. Qiao, J. H. Wang, G. Wagner, and E. L. Reinherz. HIV-1 broadly neutralizing antibody extracts its epitope from a kinked gp41 ectodomain region on the viral membrane. *Immunity*, 28(1):52–63, January 2008.

[61] Rosa M F Cardoso, Florence M Brunel, Sharon Ferguson, Michael Zwick, Dennis R Burton, Philip E Dawson, and Ian A Wilson. Structural basis of enhanced binding of extended and helically constrained peptide epitopes of the broadly neutralizing HIV-1 antibody 4E10. *Journal of Molecular Biology*, 365(5):1533–44, February 2007. PMID: 17125793.

[62] R. M. Cardoso, M. B. Zwick, R. L. Stanfield, R. Kunert, J. M. Binley, H. Katinger, D. R. Burton, and I. A. Wilson. Broadly neutralizing anti-HIV antibody 4E10 recognizes a helical conformation of a highly conserved fusion-associated motif in gp41. *Immunity*, 22(2):163–73, February 2005.

[63] E. M. Scherer, D. P. Leaman, M. B. Zwick, A. J. McMichael, and D. R. Burton. Aromatic residues at the edge of the antibody combining site facilitate viral glycoprotein recognition through membrane interactions. *Proc Natl Acad Sci U S A*, 107(4):1529–34, January 2010.

[64] H. Xu, L. Song, M. Kim, M. A. Holmes, Z. Kraft, G. Sellhorn, E. L. Reinherz, L. Stamatatos, and R. K. Strong. Interactions between lipids and human Anti-HIV antibody 4E10 can be reduced without ablating neutralizing activity. *J Virol*, 84(2):1076–88, November 2009.

[65] Y. Li, K. Svehla, M. K. Louder, D. Wycuff, S. Phogat, M. Tang, S. A. Migueles, X. Wu, A. Phogat, G. M. Shaw, M. Connors, J. Hoxie, J. R. Mascola, and R. Wyatt. Analysis of neutralization specificities in polyclonal sera derived from human immunodeficiency virus type 1-infected individuals. *J Virol*, 83(2):1045–59, January 2009.

[66] D. N. Sather, J. Armann, L. K. Ching, A. Mavrantoni, G. Sellhorn, Z. Caldwell, X. Yu, B. Wood, S. Self, S. Kalams, and L. Stamatatos. Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. *J Virol*, 83(2):757–69, January 2009.

[67] N. R. Derby, Z. Kraft, E. Kan, E. T. Crooks, S. W. Barnett, I. K. Srivastava, J. M. Binley, and L. Stamatatos. Antibody responses elicited in macaques immunized with human immunodeficiency virus type 1 (HIV-1) SF162-derived gp140 envelope immunogens: comparison with those elicited during homologous simian/human immunodeficiency virus SHIVSF162P4 and heterologous HIV-1 infection. *J Virol*, 80(17):8745–62, September 2006.

[68] M. Kim, Z. Qiao, J. Yu, D. Montefiori, and E. L. Reinherz. Immunogenicity of recombinant human immunodeficiency virus type 1-like particles expressing gp41 derivatives in a pre-fusion state. *Vaccine*, 25(27):5102–14, June 2007.

[69] M. Law, T. Maruyama, J. Lewis, E. Giang, A. W. Tarr, Z. Stamataki, P. Gastaminza, F. V. Chisari, I. M. Jones, R. I. Fox, J. K. Ball, J. A. McKeating, N. M. Kneteman, and D. R. Burton. Broadly neutralizing antibodies protect against hepatitis c virus quasispecies challenge. *Nat Med*, 14(1):25–7, January 2008.

[70] S Phogat, K Svehla, M Tang, A Spadaccini, J Muller, J Mascola, I Berkower, and R Wyatt. Analysis of the human immunodeficiency virus type 1 gp41 membrane proximal external region arrayed on hepatitis b surface antigen particles. *Virology*, 373(1):72–84, March 2008. PMID: 18155743.

[71] Douglas S Watson and Francis C Szoka. Role of lipid structure in the humoral immune response in mice to covalent lipid-peptides from the membrane proximal region of HIV-1 gp41. *Vaccine*, 27(34):4672–4683, July 2009. PMID: 19520200.

[72] D. A. Calarese, C. N. Scanlan, M. B. Zwick, S. Deechongkit, Y. Mimura, R. Kunert, P. Zhu, M. R. Wormald, R. L. Stanfield, K. H. Roux, J. W. Kelly, P. M. Rudd, R. A. Dwek, H. Katinger, D. R. Burton, and I. A. Wilson. Antibody domain exchange is an immunological solution to carbohydrate cluster recognition. *Science*, 300(5628):2065–71, June 2003.

[73] G. Ofek, M. Tang, A. Sambor, H. Katinger, J. R. Mascola, R. Wyatt, and P. D. Kwong. Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope. *J Virol*, 78(19):10724–37, October 2004.

[74] T. Zhou, L. Xu, B. Dey, A. J. Hessell, D. Van Ryk, S. H. Xiang, X. Yang, M. Y. Zhang, M. B. Zwick, J. Arthos, D. R. Burton, D. S. Dimitrov, J. Sodroski, R. Wyatt, G. J. Nabel, and P. D. Kwong. Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature*, 445(7129):732–7, February 2007.

[75] D. C. Ekiert, G. Bhabha, M. A. Elsliger, R. H. Friesen, M. Jongeneelen, M. Throsby, J. Goudsmit, and I. A. Wilson. Antibody recognition of a highly conserved influenza virus epitope. *Science*, 324(5924):246–51, April 2009.

[76] Philip R. Dormitzer, Jeffrey B. Ulmer, and Rino Rappuoli. Structure-based antigen design: a strategy for next generation vaccines. *Trends in Biotechnology*, 26(12):659–667, December 2008.

[77] Rhiju Das and David Baker. Macromolecular modeling with rosetta. *Annual Review of Biochemistry*, 77:363–382, 2008. PMID: 18410248.

[78] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*, 331(1):281–99, August 2003.

[79] F. W. Studier. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif*, 41(1):207–34, May 2005.

[80] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read. Phaser crystallographic software. *J Appl Crystallogr*, 40(Pt 4):658–674, August 2007.

[81] A. Vagin and A. Teplyakov. MOLREP: an automated program for molecular replacement. *J Appl Crystallogr*, 30, 1997.

[82] E. Potterton, P. Briggs, M. Turkenburg, and E. Dodson. A graphical user interface to the CCP4 program suite. *Acta Crystallogr D Biol Crystallogr*, 59(Pt 7):1131–7, July 2003.

[83] P. Emsley and K. Cowtan. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*, 60(Pt 12 Pt 1):2126–32, December 2004.

[84] G. N. Murshudov, A. A. Vagin, and E. J. Dodson. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr*, 53(Pt 3):240–55, May 1997.

[85] J. Painter and E. A. Merritt. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr D Biol Crystallogr*, 62(Pt 4):439–50, April 2006.

[86] J. Painter and E. A. Merritt. TLSMD web server for the generation of multi-group TLS models. *J Appl Crystallogr*, 39:109–111, 2006.

[87] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr*, 26, 1993.

[88] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, 3rd Arendall W. B., J. Snoeyink, J. S. Richardson, and D. C. Richardson. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res*, 35(Web Server issue):W375–83, July 2007.

[89] M. Law, R. M. Cardoso, I. A. Wilson, and D. R. Burton. Antigenic and immunogenic study of membrane-proximal external region-grafted gp120 antigens by a DNA prime-protein boost immunization strategy. *J Virol*, 81(8):4272–85, April 2007.

[90] F. M. Brunel, M. B. Zwick, R. M. Cardoso, J. D. Nelson, I. A. Wilson, D. R. Burton, and P. E. Dawson. Structure-function analysis of the epitope for 4E10, a broadly neutralizing human immunodeficiency virus type 1 antibody. *J Virol*, 80(4):1680–7, February 2006.

[91] J. S. Klein, P. N. Gnanapragasam, R. P. Galimidi, C. P. Foglesong, Jr. West A. P., and P. J. Bjorkman. Examination of the contributions of size and avidity to the neutralization mechanisms of the anti-HIV antibodies b12 and 4E10. *Proc Natl Acad Sci U S A*, April 2009.

[92] S. J. Hubbard and J. M. Thornton. NACCESS. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, 1993.

[93] W. L. DeLano. The PyMOL molecular graphics system. *DeLano Scientific, Palo Alto, CA, USA*, 2002.

[94] D. Opalka, C. E. Lachman, S. A. MacMullen, K. U. Jansen, J. F. Smith, N. Chirmule, and M. T. Esser. Simultaneous quantitation of antibodies to neutralizing epitopes on virus-like particles for human papillomavirus types 6, 11, 16, and 18 by a multiplexed luminex assay. *Clin Diagn Lab Immunol*, 10(1):108–15, January 2003.

[95] S. Munir Alam, Marco Morelli, S. Moses Dennison, Hua-Xin Liao, Ruijun Zhang, Shi-Mao Xia, Sophia Rits-Volloch, Li Sun, Stephen C. Harrison, Barton F. Haynes, and Bing Chen. Role of HIV membrane in neutralization by two broadly neutralizing antibodies. *Proceedings of the National Academy of Sciences*, 106(48):20234–20239, December 2009.

[96] B I Dahiyat, D B Gordon, and S L Mayo. Automated design of the surface positions of protein helices. *Protein Science: A Publication of the Protein Society*, 6(6):1333–7, June 1997. PMID: 9194194.

[97] S. M Malakauskas and S. L Mayo. Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol*, 5(6):470–5, 1998.

[98] Gautam Dantas, Brian Kuhlman, David Callender, Michelle Wong, and David Baker. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *Journal of Molecular Biology*, 332(2):449–460, September 2003. PMID: 12948494.

[99] Aaron Korkegian, Margaret E Black, David Baker, and Barry L Stoddard. Computational thermostabilization of an enzyme. *Science (New York, N.Y.)*, 308(5723):857–860, May 2005. PMID: 15879217.

[100] Michael S. Lawrence, Kevin J. Phillips, and David R. Liu. Supercharging proteins can impart unusual resilience. *Journal of the American Chemical Society*, 129(33):10110–10112, 2007.

[101] Loren L Looger, Mary A Dwyer, James J Smith, and Homme W Hellinga. Computational design of receptor and sensor proteins with novel functions. *Nature*, 423(6936):185–190, May 2003. PMID: 12736688.

[102] Tanja Kortemme and David Baker. Computational design of protein-protein interactions. *Current Opinion in Chemical Biology*, 8(1):91–97, February 2004. PMID: 15036162.

[103] Eliyahu Yosef, Regina Politi, Mee H Choi, and Julia M Shifman. Computational design of calmodulin mutants with up to 900-fold increase in binding specificity. *Journal of Molecular Biology*, 385(5):1470–1480, February 2009. PMID: 18845160.

[104] Gevorg Grigoryan, Aaron W. Reinke, and Amy E. Keating. Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature*, 458(7240):859–864, April 2009.

[105] Shaun M Lippow, K Dane Wittrup, and Bruce Tidor. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nature Biotechnology*, 25(10):1171–1176, October 2007. PMID: 17891135.

[106] Jaafar N Haidar, Brian Pierce, Yong Yu, Weiwei Tong, Michael Li, and Zhiping Weng. Structure-based design of a t-cell receptor leads to nearly 100-fold improvement in binding affinity for pepMHC. *Proteins*, 74(4):948–960, March 2009. PMID: 18767161.

[107] Daniel J Mandell and Tanja Kortemme. Computer-aided design of functional protein interactions. *Nat Chem Biol*, 5(11):797–807, November 2009.

[108] A Su and S L Mayo. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Science: A Publication of the Protein Society*, 6(8):1701–1707, August 1997. PMID: 9260282.

[109] Xueling Wu, Zhi-Yong Yang, Yuxing Li, Carl-Magnus Hogerkorp, William R Schief, Michael S Seaman, Tongqing Zhou, Stephen D Schmidt, Lan Wu, Ling Xu, Nancy S Longo, Krisha McKee, Sijy O'Dell, Mark K Louder, Diane L Wycuff, Yu Feng, Martha Nason, Nicole Doria-Rose, Mark Connors, Peter D Kwong, Mario Roederer, Richard T Wyatt, Gary J Nabel, and John R Mascola. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science (New York, N.Y.)*, 329(5993):856–861, August 2010. PMID: 20616233.

[110] Adrian A Canutescu and Roland L Dunbrack. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science: A Publication of the Protein Society*, 12(5):963–72, May 2003. PMID: 12717019.

[111] Carol A Rohl, Charlie E M Strauss, Kira M S Misura, and David Baker. Protein structure prediction using rosetta. *Methods in Enzymology*, 383:66–93, 2004. PMID: 15063647.

[112] Will Sheffler and David Baker. RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Science: A Publication of the Protein Society*, 18(1):229–239, January 2009. PMID: 19177366.

[113] M Selmer, S Al-Karadaghi, G Hirokawa, A Kaji, and A Liljas. Crystal structure of thermotoga maritima ribosome recycling factor: a tRNA mimic. *Science (New York, N.Y.)*, 286(5448):2349–2352, December 1999. PMID: 10600747.

[114] Ian A. Wilson and Robyn L. Stanfield. Antibody-antigen interactions. *Current Opinion in Structural Biology*, 3(1):113–118, February 1993.

[115] Jerry Tsai, Richard Bonneau, Alexandre V Morozov, Brian Kuhlman, Carol A Rohl, and David Baker. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, 53(1):76–87, October 2003. PMID: 12945051.

[116] P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 393(6686):648–59, June 1998.

[117] Hiroaki Nakano, Takuya Yoshida, Susumu Uchiyama, Masako Kawachi, Hitomi Matsuo, Takayuki Kato, Atsushi Ohshima, Yoshiharu Yamaichi, Takeshi Honda, Hiroaki Kato, Yuriko Yamagata, Tadayasu Ohkubo, and Yuji Kobayashi. Structure and binding mode of a ribosome recycling factor (RRF) from mesophilic bacterium. *The Journal of Biological Chemistry*, 278(5):3427–3436, January 2003. PMID: 12411440.

[118] Daniel N Wilson, Frank Schluenzen, Joerg M Harms, Takuya Yoshida, Tadayasu Ohkubo, Renate Albrecht, Joerg Buerger, Yuji Kobayashi, and Paola Fucini. X-ray crystallography study on ribosome recycling: the mechanism of binding and action of RRF on the 50S ribosomal subunit. *The EMBO Journal*, 24(2):251–260, January 2005. PMID: 15616575.

[119] Daniel J Mandell and Tanja Kortemme. Backbone flexibility in computational protein design. *Current Opinion in Biotechnology*, 20(4):420–428, August 2009. PMID: 19709874.

[120] S Annie Gai and K Dane Wittrup. Yeast surface display for protein engineering and characterization. *Current Opinion in Structural Biology*, 17(4):467–473, August 2007. PMID: 17870469.

[121] Benjamin J Hackel, Atul Kapila, and K Dane Wittrup. Picomolar affinity fibronectin domains engineered utilizing loop length diversity, recursive mutagenesis, and loop shuffling. *Journal of Molecular Biology*, 381(5):1238–1252, September 2008. PMID: 18602401.

[122] Philip A Romero and Frances H Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews. Molecular Cell Biology*, 10(12):866–876, December 2009. PMID: 19935669.

[123] Robert J Hayes, Jorg Bentzien, Marie L Ary, Marian Y Hwang, Jonathan M Jacinto, Jst Vielmetter, Anirban Kundu, and Bassil I Dahiyat. Combining computational and experimental screening for rapid optimization of protein properties. *Proceedings of the National Academy of Sciences of the United States of America*, 99(25):15926–15931, December 2002. PMID: 12446841.

[124] Thomas P Treynor, Christina L Vizcarra, Daniel Nedelcu, and Stephen L Mayo. Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):48–53, January 2007. PMID: 17179210.

[125] C F Barbas, E Bjrling, F Chiodi, N Dunlop, D Cababa, T M Jones, S L Zebedee, M A Persson, P L Nara, and E Norrby. Recombinant human fab fragments neutralize human type 1 immunodeficiency virus in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 89(19):9339–9343, October 1992. PMID: 1384050.

[126] Dennis R Burton and Robin A Weiss. AIDS/HIV. a boost for HIV vaccine design. *Science (New York, N.Y.)*, 329(5993):770–773, August 2010. PMID: 20705840.

[127] J R Mascola, G Stiegler, T C VanCott, H Katinger, C B Carpenter, C E Hanson, H Beary, D Hayes, S S Frankel, D L Birx, and M G Lewis. Protection of macaques against vaginal transmission of a pathogenic HIV-1/SIV chimeric virus by passive infusion of neutralizing antibodies. *Nature Medicine*, 6(2):207–210, February 2000. PMID: 10655111.

[128] P W Parren, P A Marx, A J Hessell, A Luckay, J Harouse, C Cheng-Mayer, J P Moore, and D R Burton. Antibody protects macaques against vaginal challenge with a pathogenic r5 simian/human immunodeficiency virus at serum levels giving complete neutralization in vitro. *Journal of Virology*, 75(17):8340–8347, September 2001. PMID: 11483779.

[129] Gunilla B. Karlsson Hedestam, Ron A.M. Fouchier, Sanjay Phogat, Dennis R. Burton, Joseph Sodroski, and Richard T. Wyatt. The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus. *Nat Rev Micro*, 6(2):143–155, February 2008.

[130] Lei Chen, Young Do Kwon, Tongqing Zhou, Xueling Wu, Sijy O'Dell, Lisa Cavacini, Ann J. Hessell, Marie Pancera, Min Tang, Ling Xu, Zhi-Yong Yang, Mei-Yun Zhang, James Arthos, Dennis R. Burton, Dimiter S. Dimitrov, Gary J. Nabel, Marshall R. Posner, Joseph Sodroski, Richard Wyatt, John R. Mascola, and Peter D. Kwong. Structural basis of immune evasion at the site of CD4 attachment on HIV-1 gp120. *Science*, 326(5956):1123–1127, November 2009.

[131] Tongqing Zhou, Ivelin Georgiev, Xueling Wu, Zhi-Yong Yang, Kaifan Dai, Andrs Finzi, Young Do Kwon, Johannes F Scheid, Wei Shi, Ling Xu, Yongping Yang, Jiang Zhu, Michel C Nussenzweig, Joseph Sodroski, Lawrence Shapiro, Gary J Nabel, John R Mascola, and Peter D Kwong. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science (New York, N.Y.)*, 329(5993):811–817, August 2010. PMID: 20616231.

[132] Ginger Chao, Wai L Lau, Benjamin J Hackel, Stephen L Sazinsky, Shaun M Lippow, and K Dane Wittrup. Isolating and engineering human antibodies using yeast surface display. *Nature Protocols*, 1(2):755–768, 2006. PMID: 17406305.

[133] T A Kunkel. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proceedings of the National Academy of Sciences of the United States of America*, 82(2):488–492, January 1985. PMID: 3881765.

[134] T A Kunkel, K Bebenek, and J McClary. Efficient site-directed mutagenesis using uracil-containing DNA. *Methods in Enzymology*, 204:125–139, 1991. PMID: 1943776.

[135] Y. Li, S. A. Migueles, B. Welcher, K. Svehla, A. Phogat, M. K. Louder, X. Wu, G. M. Shaw, M. Connors, R. T. Wyatt, and J. R. Mascola. Broad HIV-1 neutralization mediated by CD4-binding site antibodies. *Nat Med*, 13(9):1032–4, September 2007.

[136] G. Grigoryan and A. E Keating. Structural specicity in coiled-coil interactions. *Curr Opin Struct Biol*, (0), 2008.

[137] Irene M A Nooren and Janet M Thornton. Diversity of protein-protein interactions. *The EMBO Journal*, 22(14):3486–3492, July 2003. PMID: 12853464.

[138] Emmanuel D Levy and Jose B Pereira-Leal. Evolution and dynamics of protein interactions and networks. *Current Opinion in Structural Biology*, 18(3):349–357, June 2008. PMID: 18448325.

[139] James J Havranek and David Baker. Motif-directed flexible backbone design of functional interactions. *Protein Science: A Publication of the Protein Society*, 18(6):1293–1305, June 2009. PMID: 19472357.

[140] Jason S McLellan, Man Chen, Albert Kim, Yongping Yang, Barney S Graham, and Peter D Kwong. Structural basis of respiratory syncytial virus neutralization by motavizumab. *Nature Structural & Molecular Biology*, 17(2):248–250, February 2010. PMID: 20098425.

[141] Herren Wu, David S Pfarr, Syd Johnson, Yambasu A Brewah, Robert M Woods, Nita K Patel, Wendy I White, James F Young, and Peter A Kiener. Development of motavizumab, an ultra-potent antibody for the prevention of respiratory syncytial virus infection in the upper and lower respiratory tract. *Journal of Molecular Biology*, 368(3):652–665, May 2007. PMID: 17362988.

[142] R. B Hill, D. P Raleigh, A. Lombardi, and W. F DeGrado. De novo design of helical bundles as models for understanding protein folding and function. *Acc Chem Res*, 33(11):745–54, 2000.

[143] S. T Walsh, A. L Lee, W. F DeGrado, and A. J Wand. Dynamics of a de novo designed three-helix bundle protein studied by 15N, 13C, and 2H NMR relaxation methods. *Biochemistry*, 40(32):9560–9, 2001.

[144] C E Schafmeister, S L LaPorte, L J Miercke, and R M Stroud. A designed four helix bundle protein with native-like structure. *Nature Structural Biology*, 4(12):1039–1046, December 1997. PMID: 9406555.

[145] Jennifer R Calhoun, Hidetoshi Kono, Steven Lahr, Wei Wang, William F DeGrado, and Jeffery G Saven. Computational design and characterization of a monomeric helical dinuclear metalloprotein. *Journal of Molecular Biology*, 334(5):1101–1115, December 2003. PMID: 14643669.

[146] J W Bryson, J R Desjarlais, T M Handel, and W F DeGrado. From coiled coils to small globular proteins: design of a native-like three-helix bundle. *Protein Science: A Publication of the Protein Society*, 7(6):1404–1414, June 1998. PMID: 9655345.

[147] Philip Bradley and David Baker. Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. *Proteins*, 65(4):922–929, December 2006. PMID: 17034045.

[148] Carol A Rohl. Protein structure estimation from minimal restraints using rosetta. *Methods in Enzymology*, 394:244–260, 2005. PMID: 15808223.

[149] M Bansal, S Kumar, and R Velavan. HELANAL: a program to characterize helix geometry in proteins. *Journal of Biomolecular Structure & Dynamics*, 17(5):811–819, April 2000. PMID: 10798526.

[150] Till Rehm, Robert Huber, and Tad A Holak. Application of NMR in structural proteomics: screening for proteins amenable to structural analysis. *Structure (London, England: 1993)*, 10(12):1613–1618, December 2002. PMID: 12467568.

[151] Xiaodong Xiao, Weizao Chen, Yang Feng, Zhongyu Zhu, Ponraj Prabakaran, Yanping Wang, Mei-Yun Zhang, Nancy S Longo, and Dimiter S Dimitrov. Germline-like predecessors of broadly neutralizing antibodies lack measurable binding to HIV-1 envelope glycoproteins: implications for evasion of immune responses and design of vaccine immunogens. *Biochemical and Biophysical Research Communications*, 390(3):404–409, December 2009. PMID: 19748484.

[152] M Margarida Souto-Carneiro, Nancy S Longo, Daniel E Russ, Hong wei Sun, and Peter E Lipsky. Characterization of the human ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *Journal of Immunology (Baltimore, Md.: 1950)*, 172(11):6790–6802, June 2004. PMID: 15153497.

[153] Elisabeth L Humphris and Tanja Kortemme. Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design. *Structure (London, England: 1993)*, 16(12):1777–1788, December 2008. PMID: 19081054.